

Measurement
in
Today's Schools
THIRD EDITION

Measurement in Today's Schools

by
C. C. Ross

Revised by
Julian C. Stanley

PROFESSOR OF EDUCATION
UNIVERSITY OF WISCONSIN

THIRD EDITION

MLSU - CENTRAL LIBRARY



19913EX

Englewood Cliffs, N. J.
PRENTICE-HALL, INC.

COPYRIGHT, 1941, 1947, 1954, BY
PRENTICE-HALL, INC
ENGLEWOOD CLIFFS, N. J.

All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publishers.

L. C. Cat. Card No. 54-8846

<i>First printing</i>	<i>May, 1954</i>
<i>Second printing</i>	<i>February, 1955</i>
<i>Third printing</i>	<i>March, 1956</i>
<i>Fourth printing</i>	<i>April, 1958</i>
<i>Fifth printing</i>	<i>January, 1959</i>
<i>Sixth printing</i>	<i>February, 1960</i>
<i>Seventh printing</i>	<i>February, 1961</i>
<i>Eighth printing</i>	<i>January, 1962</i>

T:5

J4



19913

Preface to the Third Edition

In this third edition of Professor Ross's book, the general framework of the earlier editions has been retained but deletions, insertions, or other changes have been made on nearly every page to modernize the content and to increase readability.

"The Statistical Analysis of Test Results," Chapter 8 in previous editions, has been rewritten almost completely and moved forward to become Chapter 3 and thus furnish needed background early in the book. Fifty five-option multiple choice instructional items in Appendix A, with answers and explanations in Appendix E, and the review of square root computation in Appendix D were added to round out this material.

The old Chapter 12, "Practice," has been absorbed by Chapter 11 now called "Motivation and Practice as Related to Testing." The chapter on "School Marks" was dropped since it overlapped other portions of the book and is now outdated.

Chapter 17, "Some Present Trends," and all six appendices are completely new. Appendix B, "A Simplified Item Analysis Procedure," is based upon hitherto unpublished studies by the writer. It sets forth, perhaps for the first time in an American elementary measurements book, a simple, illustrated, complete technique for determining the discriminating power and difficulty of each question in a test and several characteristics of the test scores. Appendix C, "Scoring Rearrangement (Ranking) Test Items," contains a table with values preferable to those commonly used for this type of item. Table F, "Publishers of Standardized Tests," replaces a shorter list formerly placed at the end of the final chapter.

Three other additions should make the book easier to use: chapter sub-headings in the Table of Contents, a List of Tables, and an Author Index.

I am indebted to a number of persons for assistance during the course of this revision. Mr. Gordon D. Mock checked many bibliographic entries. Professor Chester W. Harris supplied excellent comments concerning the items in Appendix A, Professor Robert L. Ebel made several helpful suggestions with reference to Appendices B and C, and Mrs. Margaret T. Aldridge and Professor Eric F. Gardner reviewed Chapter 3.

In particular, I am grateful to my wife, Rose S. Stanley, for her painstaking secretarial work and for otherwise expediting the revision.

JULIAN C. STANLEY

Preface to the Second Edition

Since publication of the first edition of this book, experimentation in the field of measurement has made considerable progress. In this revision, the author has taken advantage of developments in this field, revising almost all the material from the first edition and including a great deal of altogether new material. All bibliographies and citations have been revised, a list of leading publishers of tests has been added to the final chapter.

Chapter tests and exercises incorporated directly into each chapter of the first edition, have now been compiled into a separate workbook. This arrangement is designed to save valuable time for the student working on the exercises and for the teacher correcting them.

Sincere appreciation is due Mrs. Billy Whitlow Smith for considerable work both in the preparation of the manuscript and in correcting proof. Her assistance has been invaluable at every stage of the book's progress.

Preface to the First Edition

It is doubtless true that more progress in measurement has been made during the past quarter of a century than during all the years preceding. But the pattern of the measurement books has remained much the same. They have been very definitely centered about *subject matter*. The treatment has usually been organized around the conventional school subjects, *and much space has been devoted to lists and descriptions of the measuring instruments available*.

Authors of these texts have encountered obstacles increasingly difficult to surmount. The rapid increase in the number of tests and scales published has made it impossible to keep the books either complete or up to date. Even the most carefully compiled list of selected tests was likely to be rendered obsolete by the publication of better tests before the book was off the press. Fortunately, in recent years the appearance of rather complete and frequently revised bibliographies of published tests, together with critical evaluations, has made detailed lists and descriptions of available measuring instruments in textbooks no longer necessary.

Meanwhile instructors in measurement have manifested a growing dissatisfaction with existing texts on the subject. For example the typical class in measurement for high school teachers has consisted of persons representing a variety of fields, but no one person has been interested in more than two or three of those discussed in the textbook, the rest of the material being largely deadwood. At the same time the enormous expansion of the experimental literature relating to measurement has had to be considered in any course that is at all adequate. And here the average book has left much to be desired.

Fifteen years' experience in teaching educational measurement to college classes has led the author to attempt a functional approach to the subject. The present work is the outgrowth of this experience. The emphasis is therefore, not so much upon the description of the tools themselves as upon the multitude of problems relating to their intelligent use and interpretation by classroom teachers and school administrators.

It appears to the author that the time has come for a critical appraisal of measurement in today's schools and for a careful search for generaliza-

tions to guide both theory and practice. The experimental evidence supporting these generalizations has been examined, and wherever possible reported in the language of the original author.

Since the functions of measurement are much the same on all educational levels, the illustrations have been drawn from both the elementary school and the secondary school, and to some extent from college. It is hoped that the book will be found useful to teachers, and to prospective teachers, regardless of the subject or the level of instruction.

In the preparation of the book the author has incurred obligations that are numerous and great. His first major indebtedness has been to his former teachers, notably Professors Edward L. Thorndike and William A. McCall, of Teachers College, Columbia University. The heavy obligation which the author owes to his co-workers in the field of measurement, upon whose publications he has freely drawn, is indicated by the numerous citations throughout the book. The fullest co-operation of these authors and their publishers is gratefully acknowledged. Special thanks are due to Professor A. B. Crawford, who has used a preliminary edition of the book at Transylvania College and at the University of Kentucky, and who has made numerous constructive suggestions, and to Professor G. M. Ruch, of the United States Office of Education, who has read the manuscript, and whose pertinent criticisms have been invaluable. Finally, the author is indebted to his own students, who for three years have used a preliminary edition of the book and who have offered many suggestions that have contributed greatly to its improvement.

C. C. Ross

Table of Contents

Part I

THE PROBLEM OF MEASUREMENT

CHAPTER

PAGE

1. MEASUREMENT IN THE MODERN WORLD 3
Measurement in science, 4 Measurement in education, 13
2. THE HISTORICAL DEVELOPMENT OF MEASUREMENT IN EDUCATION 27
Introduction, 27 The history of intelligence tests, 30 The history of achievement tests 38 The history of character, personality, and interest measurement 46 Some important publications 52 Some relatively recent tendencies, 56
- 3 THE STATISTICAL ANALYSIS OF TEST RESULTS 60
General considerations, 60 Classification and tabulation, 61 Some elementary notions concerning quantitative data 69 Finding the mode the median, and the mean, 75 Measures of variability or scatter 81 Measures of relationship, 85 Measures of error, 101 Summary, 104 Instructional test items, 104
4. THE CHARACTERISTICS OF A SATISFACTORY MEASURING INSTRUMENT 106
Introduction, 106 Validity, 107 Reliability, 121 Usability, 127 Some generalizations regarding the problem of measurement, 131

Part II

THE CONSTRUCTION OF TEACHER-MADE TESTS

- 5 GENERAL PRINCIPLES OF TEST CONSTRUCTION 139
Planning the test, 140 Preparing the test, 147 Trying out the test 155 Evaluating the test, 159
- 6 PRINCIPLES OF CONSTRUCTING SPECIFIC TYPES OF OBJECTIVE TESTS 163
Introduction, 163 Simple-recall tests, 167 Completion tests, 170 Alternative-response tests 174 Multiple-choice tests, 179 Matching tests 186 Rearrangement tests, 190

7. THE CONSTRUCTION AND USE OF ESSAY EXAMINATIONS 192

Limitations of the essay examination 193 Advantages of the essay examination, 196 Suggestions for improving essay examinations 197

Part III

THE TESTING PROGRAM

8 STEPS IN THE TESTING PROGRAM 209

Determining the purpose of the program 212 Selecting the appropriate test or tests 214 Administering the tests 225 Scoring the tests 230 Analyzing and interpreting the scores, 234 Applying the results, 235 Retesting to determine the success of the program, 236 Making suitable records and reports, 236

9 THE GRAPHICAL REPRESENTATION OF EDUCATIONAL DATA 247

The value of graphs 247 Representing the record of an individual, 254 Representing a frequency distribution 258 Representing two or more distributions, 261. General suggestions for constructing graphs, 271

10 THE USES AND LIMITATIONS OF NORMS 274

Norms and Standards, 274 Raw scores and derived scores 276 The use of norms in interpreting scores on intelligence tests 279 The use of norms in interpreting scores on achievement tests 290 Methods of comparing intelligence and achievement, 296 The use of norms in interpreting scores on personality tests, 299

Part IV

MEASUREMENT IN INSTRUCTION

11 MOTIVATION AND PRACTICE AS RELATED TO TESTING 303

The problem of motivation 303 The relation of measurement to motivation in teaching 304 The relation of measurement to motivation in learning, 306 Some educational implications of motivation studies, 325 Practice effect 326

12 DIAGNOSIS 328

The problem of diagnosis in education 328 The techniques of diagnosis 332

13 CLASSIFICATION AND PROMOTION 317

The nature and educational significance of human variability, 347 The activity movement, 356 Homogeneous or ability groups 357

14 EVALUATION IN GUIDANCE 367

The meaning and importance of guidance, 367 The place of measurement in guidance, 369 Guidance is a co-operative venture, 370

15 EVALUATION OF SCHOOLS 373

The problem of evaluation 373 General principles of evaluation, 380 Evaluating various aspects of the school 383

CHAPTER

	PAGE
16. PUBLIC RELATIONS . . .	400
The problem, 400 Ordinary agencies of public information, 402 Official publications, 404 Report cards and letters to parents, 406 Other avenues of public information, 413 Mobilizing public opinion, 414	
17. SOME PRESENT TRENDS	416

APPENDICES

APPENDIX

A. FIFTY QUESTIONS TO HELP YOU LEARN STATISTICS .	429
B. A SIMPLIFIED ITEM-ANALYSIS PROCEDURE	436
Preparing the items, 436 A measure of discrimination, 437 A measure of difficulty, 440 An illustrative analysis, 440 A discrimination table, 447 Obtaining the mean and the standard deviation, 452 A simplified procedure for obtaining a reliability coefficient, 452	
C. SCORING REARRANGEMENT (RANKING) TEST ITEMS	454
D. THE COMPUTATION OF SQUARE ROOTS . . .	456
E. ANSWERS TO QUESTIONS IN APPENDIX A . . .	459
F. PUBLISHERS OF STANDARDIZED TESTS	464
AUTHOR INDEX	467
SUBJECT INDEX	473

List of Tables

TABLE	PAGE
1 The Estimated Grade-Value and Percentage Marks Assigned to an English Composition by One Hundred Teachers	40
2 Percentage Values Assigned to Ten Essay Examination Papers by Twenty Four Examiners	42
3 A Class Record for a Reading Readiness Test	62
4 Reading Readiness Scores from Table 3 Arranged in Order of Size and Rank Order and Tabulated	63
5 An Illustration of the Process of Making a Grouped Frequency Distribution	65
6 Distribution of Reading Readiness Scores for Six Schools in a Certain City	66
7 The Chronological Educational and Mental Ages of the 20 Pupils in an Eighth-Grade Class	67
8 A Two-Way Distribution of Mental Age and Educational Age for an Eighth Grade Class	68
9 An Extremely Simple Scoring Method	70
10 Five Consecutive Exploratory Steps in Assigning Overall Scores to 31 English Themes	70
11 Dividing the Four <i>D</i> s from the 11 Category Distribution of Table 10 into Three Parts	74
12 The Process of Locating the Median	76
13 A Short Way to Compute the Mean	80
14 The Process of Computing the Quartile Deviation <i>Q</i>	82
15 A Simplified Way to Compute the Standard Deviation	84
16 A Scatter Diagram Illustrating Negative Correlation Between Chronological Age and Educational Age for 20 Eighth Graders	87
17 Pretest and Midterm Scores of 43 Graduate Students on Two Teacher Made Objective Tests in Intermediate Statistics	91
18 Computation of the <i>r</i> Between Scores on Two Teacher Made Tests	92
19 The Computation of Rho for Each of Two Students Who Arranged Six Historical Events in Chronological Order	93
20 The Various Values of Rho (ρ) for All Possible Sums of Squared Deviations (ΣD^2) for <i>N</i> s from 2 through 10	96
21 Estimating the Coefficient of Correlation by the Spearman Rank Difference Method	99
22 A Simple Expectancy Table Based upon the 43 Pairs of Scores in the Table 18 Scattergram, for Which $r = .53$	100
23 Effects of Constant and Variable Errors on Certain Types of Statistics	103
24 Intercorrelations of Intelligence Test Scores and Five-Semester Average Grades for 284 Seniors (124 Boys 160 Girls) in a Los Angeles High School	110
25 Rankings of Test Items According to Frequency of Use as Revealed by Two Studies	164
26 Plan for a Testing Program for the Elementary School	210
27 Advantages and Limitations of Standardized and Nonstandardized Tests of Achievement	216 217

LIST OF TABLES

xiii

TABLE	PAGE
28 Classification of Tests in <i>The Fourth Mental Measurements Yearbook</i> (1933)	218
29 Otis Scale for Rating Standard Tests	220
30 Cole-Von Borgerstrode Scale for Rating Standardized Tests	222
31 A Table for Computing Months Since Last Birthday	283
32 Table for Equating Intelligence Quotient Values	286
33 Point Standing for the First and Second Semesters for Low Ranking Fresh men Who Were Told Their Intelligence Test Scores as Compared with Those Who Were Not	320
34 Distribution of Spelling Difficulties and Successful Remedies	342
35 Frequencies with Which Various Provisions for Individual Differences Were Reported in Use or in Use with Unusual Success	354
36 Trends Toward Greater Provisions for Individual Differences in Elementary Schools	355
37 The Main Methods of Evaluation Used by the Co-operative Study of Secondary School Standards with the Weight Assigned to Each	375
38 Composition of 1940 Edition of the Alpha Beta and Gamma Scales for Evaluating Secondary Schools	377
39 Use of Tests in Evaluating Schools	379
40 Summary of the Mort-Cornell Score Sheet for the Self Appraisal of School Systems	396
41 The Rank Orders of Thirteen Topics of School News According to the Interests of 5 067 School Patrons Compared with the Space Devoted to These Topics by Ten Newspapers	403
42 The 100 Items in a Five-Option Multiple-Choice Teacher Made Test Arranged According to Discriminating Power	438
43 Number of Examinees in High and Low Groups Who Chose Each Option of Item No. 30	441
44 Number of Examinees in High and Low Groups Who Chose Each Option of Item No. 90	442
45 Number of Examinees in High and Low Groups Who Chose Each Option of the 25 Least Discriminating Items	443
46 Table for Determining Whether or Not a Given Test Item Discriminates Significantly Between a High and a Low Group	448
47 Formulas for Finding ($W_L + W_H$) Values at Three Difficulty Levels	451
48 ΣD^2 Table for Scoring Rearrangement Items	455

List of Figures

FIGURE	PAGE
1 Test 7 from the Army Alpha	35
2 Test 6 from the Army Beta	36
3 A Scale for Measuring Pupils' Attitudes Toward High School	50
4 The Relative Amount of Relationship Represented by r 's of Various Sizes	89
5 IBM General-Purpose Answer Sheet	159
6 An Illustration of the Procedure Followed in Scoring Test 3 of the Terman Group Test of Mental Ability Form A	232
7 A Sample Standard Test Scoring Record	233
8 An Educational Profile for a Standardized Achievement Test	239
9 Test Data Summary from the Cumulative Guidance Record of the Department of Supervision and Curriculum Development of the National Education Association	241
10 A Cumulative Record in Graphical Form	242
11 Centile Sheet for College Men and Women on Allport-Lindzey-Vernon Study of Values	245
12 Back to School United States 1900-1953	248
13 A Rather Complex Bar Graph with High Attention Value	250-251
14 Motor Buses in Operation in the United States—Fifteen Different Charts Based upon the Same Data	252 253
15 Profile of a Pupil and the Sixth Grade Class of Which He Is a Member	255
16 The Profile of a Tenth Grade Pupil on the California Achievement Test	256
17 Profiles for a Student Tested in the Fifth and Sixth Grades	257
18 A Histogram or Column Diagram Representing the Percentage Values Assigned to an Arithmetic Paper by Forty Two Scorers	258
19 A Histogram or Column Diagram Representing the Distribution of IQ's in a Small Junior High School	259
20 A Frequency Polygon Representing the Percentage Values Assigned to an Arithmetic Paper by Forty Two Scorers	259
21 An Actual Curve Compared with the Theoretical Curve of Probability	260
22 A Percentile Curve Representing the Percentage Values Assigned to an Arithmetic Paper by Forty Two Scorers	261
23 A Percentile Curve Representing the Distribution of 83 IQ's in a Small Junior High School	261
24 Negative and Positive Skewness	262
25 Bar Graph Made on the Typewriter Showing the Distribution of 91 IQ's in a Junior High School	262
26 Bar Graph Made on the Typewriter Showing the Percentage of Pupils of Each Age Group Who Were Graduated from High School and the Percentage Who Entered High School but Did Not Graduate	263
27 Graph Made on the Typewriter Showing the Overlapping of Grades Seven, Eight, and Nine in Reading Comprehension	264

LIST OF FIGURES

vi

FIGURE	PAGE
28 Frequency Polygons Representing the Distribution of Reading Comprehension Scores on the Iowa Silent Reading Tests for the Seventh Eighth and Ninth Grades of a Certain School	260
29 Total Comprehension Scores on the Iowa Silent Reading Tests for the Seventh Eighth and Ninth Grades	266
30 The Learning of Three Groups Compared One with Full Knowledge of Progress One with Partial Knowledge of Progress and One with No Knowledge of Progress	266
31 Correct and Incorrect Location of the Norms in a Line Chart Showing Median Scores on a Reading Test	267
32 Grade Profiles for the Seventh Eighth and Ninth Grades of a Certain Junior High School Made by Connecting the Median Scores on Each Part of the Stanford Achievement Test Advanced Complete Battery Form J	268
33 A Line Graph Showing the Medians and Quartiles for Grades Four to Nine Inclusive in Reading Comprehension	269
34 The Central Tendency and Variability in Educational Age of Grades 2B to 9A Inclusive in a Small City School System	270
35 The Relation Between Standard Scores Percentile Ranks and Revised Stanford Binet IQs	290
36 The Profiles of Two Pupils Who Made the Same Total Score on a General Achievement Test	292
37 A Profile Based Upon Local Norms	297
38 The Influence of Knowledge of Progress upon Achievement in a College Class	318
39 A Study of the Influence of Praise and Reproof upon Achievement in Fourth Grade and Sixth Grade Arithmetic	321
40 The Five Levels of Educational Diagnosis	332
41 Analysis Sheet of Test 3 Metropolitan Achievement Tests Form A Arithmetic Fundamentals for a Fifth Grade Class in October	335
42 Traxler Chart of Suggested Diagnostic and Remedial Procedures in Handwriting	343 344
43 General Quality of 200 Secondary Schools as Judged by Field Committees	349
44 Distribution of Mean Scores of Seniors in Forty Nine Colleges in Pennsylvania on a Test of General Academic Knowledge	350
45 Distributions of Composite IQs on Forms L and M of the Revised Stanford Binet Intelligence Scales for a Standardization Group of 2904 Individuals of CAs 2 to 18 Years	351
46 A Suggested Technique for Evaluating the Philosophy of a Secondary School	383
47 Instructions for Using the Evaluative Criteria Developed by the Co-operative Study of Secondary School Standards	385
48 Summary of Evaluative Criteria for the Median Secondary School	386
49 An Evaluative Procedure for the Content of the Offerings in the Principal Subject-Matter Fields of a Secondary School	387
50 The Computation of Three Measures of the Adequacy of the Book Collections in the Library of a Secondary School	389
51 Evaluative Techniques for the Library Service of a Secondary School	390
52 The Computation of the Summary Score for the Guidance Service of a Secondary School	391
53 An Evaluative Technique for the Quality of Instruction in a Secondary School	393
54 A Suggested Informal Report to Parents	409
55 A Report Card Used at the University of Chicago High School	412

Measurement
in
Today's Schools
THIRD EDITION

PART I

The Problem of Measurement

I

Measurement in the Modern World

From birth to death almost every aspect of our daily lives is touched by measurement in its numerous forms. At birth the record of that important event is carefully made according to the nurse's watch. During the next few days measurements of the baby's weight and temperature are part of the daily routine of the hospital. Ever afterward, whether in school or outside, watches, clocks, balances, thermometers, money systems and other forms of measurement play prominent roles in the life of every human being.

The daily round of the typical American probably begins somewhat like this. He rises at a certain hour by the clock, bathes in water measured by the meter, and dresses in clothing of a standard size. He begins his breakfast with half a grapefruit sold by the dozen and sweetened with a pound of sugar sold by the pound. He continues with a bowl of cereal and a cup of coffee, both generously mixed with cream or milk sold by the quart. He then looks at his watch, jumps in his car, and watches the meter as he hurries to his work, for which he is paid by the hour, month, or year.

For another, year in and year out he keeps this up until he reaches the end of himself to death over a falling stock market measured by the index, rising blood pressure, expressed in points. Then the hour of his death is accurately noted, he is measured for a casket, and the date is set according to the calendar and the clock. Afterward his name is recorded in the family Bible and carved on his tombstone. His estate is figured in dollars and his widow lives the rest of her life on his income computed in per cent.

These common experiences are characteristic of the emphasis placed on measurement in the modern world. In fact, if all our various measuring devices were suddenly destroyed, contemporary civilization would collapse like a house of cards.

A. Measurement in Science

Apparently the chief problem of man has always been *adjustment*. As one writer puts it "The civilization of a race is simply the sum total of its achievements in adjusting itself to its environment."¹ The form of the problem has indeed varied somewhat from time to time, and still more has the method of meeting it. For ages the ingenuity of man was directed toward gaining practical *control* over the universe about him. At first the process was the uncritical procedure of trial and error. This fumbling way early led into such blind alleys as alchemy, astrology, and magic. Later the seers and wise men began to attempt to put together these scattered bits of experience and so in the words of Omar Khayyam, "To grasp this sorry Scheme of Things entire." Thus was born philosophy. The nature of the problem had then shifted to *understanding* the universe, rather than merely gaining control over it.

Scientific method. About three centuries ago there arose, with the experimental verification by Galileo of the laws of falling bodies, the method of modern science. Since that time man's quantitative conquest of nature has expanded not only into all branches of physics and chemistry but into biological and psychological phenomena as well. It is no exaggeration today to assert that science has revolutionized the material world in which we live. But it has done more than this, as Whitehead says, science has "practically recoloured our mentality."² As a distinguished chemist puts it "Man's *inner* and *outer necessities*, real or imagined, have made him both a Scientist and a Philosopher."³

Both the *content* and the *method* of science are important. The content of science consists of a continuously expanding body of systematized knowledge, which is the product of scientific method. The one constant and universal feature of science is its method of arriving at knowledge. John Dewey asserts that "the heart of science lies not in conclusions reached, but in the method of observation, experimentation, and mathematical reasoning by which conclusions are established."⁴

¹ Hu Shih, "The Civilization of the East and the West" in *Whither Mankind* edited by Charles A. Beard, page 27. New York: Longmans, Green & Company, 1928.

² Alfred North Whitehead, *Science and the Modern World*, page 3. New York: The Macmillan Company, 1925.

³ Richard E. Lee, *The Backgrounds and Foundations of Modern Science*, page 3. Baltimore: The Williams & Wilkins Company, 1935.

⁴ *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II*, page 450. Quoted by permission of the Society. Bloomington: Illinois Public School Publishing Company, 1938.

What, then, is the scientific method? Bertrand Russell suggests this concise formulation "The essence of the scientific method is the discovery of general laws through the study of particular facts"⁵ In another volume Russell elaborates this statement⁶

In arriving at a scientific law there are three stages the first consists in observing the significant facts, the second in arriving at a hypothesis which if it is true would account for these facts, the third in deducting from this hypothesis consequences which can be tested by observation

Conant's definition emphasizes continuity⁷

Science is an interconnected series of concepts and conceptual schemes that have developed as a result of experimentation and observation and are fruitful of further experimentation and observations In this definition the emphasis is on the word 'fruitful' Science is a speculative enterprise The validity of a new idea and the significance of a new experimental finding are to be measured by the consequences—consequences in terms of other ideas and other experiments Thus conceived science is not a quest for certainty, it is rather a quest which is successful only to the degree that it is continuous

But what is the role of measurement in scientific method? From Russell's three stage analysis above it would appear that, although measurement has but little if any bearing on the second stage in the scientific method, it is closely related to the first and third stages Measurement performs a useful function in determining what alleged facts really are facts as well as providing an exact method of describing them It is also indispensable in the final stage of testing and verification which is usually by means of specially devised experiments A critical treatise⁸ on educational measurement begins with this statement "Measurement is the principal implement of science changing that field of human endeavor from medieval gropings to a modern exactitude" The relationship is stated by Smart in the following words⁹

Of course, it must not be forgotten that our experience of sense qualities in perception serves as the basis of all scientific endeavor and that this qualitative aspect of things is in varying degrees of completeness assimilated in and through the higher categories of the several natural sciences And this assimilation is effected largely through the process of measurement which thus functions as the connecting link between mathematics and the other sciences and which is only a higher i.e. more precise and complete form of that double-sided process of comparison and discrimination which begins on the qualitative level of experience

⁵ Bertrand Russell *Science in History* Mankind op cit page 65

⁶ Bertrand Russell *The Scientific Outlook* page 57 New York W W Norton & Company Inc 1931

⁷ James Bryant Conant *Science and Common Sense* pages 25-26 New Haven Yale University Press 1931

⁸ B Othanel Smith *Logical Aspects of Educational Measurement* 182 pages New York Columbia University Press 1938

⁹ Harold R Smart *The Logic of Science* page 200 New York D Appleton and Company 1914 Used by permission of D Appleton Century Company

Brief attention will now be given to the relation of measurement to each of the principal divisions of science. The discussion will observe the conventional divisions, namely, the "pure sciences" and the "applied sciences," though we should note that in his presidential address to the American Association for the Advancement of Science, Kirtley Mather emphasized the shortcomings of these terms.¹⁰

It is significant that when scientists today are philosophizing, they are more likely to distinguish between "fundamental research" and "technological development" than between "pure science" and "applied science." The fact is, of course, that every item of scientific knowledge ever gained, in response to whatever motive, has been found sooner or later to have practical significance, either directly or indirectly, in human affairs.

Pure science is distinguished from applied science primarily on the basis of *purpose* or *motive*, and one division of pure science is distinguished from another on the basis of *subject matter*. Although the distinction is not always clear-cut, in general it may be said that pure science aims primarily at *understanding* the universe, whereas applied science aims at *predicting* and *controlling* it. In Russell's words, "Science, ever since the time of the Arabs, has had two functions: first, to enable us to *know* things and, second, to enable us to *do* things."¹¹

He also says "Science, as its name implies, is primarily knowledge, by convention it is knowledge of a certain kind. Gradually, however, the aspect of science as knowledge is being thrust into the background by the aspect of science as the power of manipulating nature."¹²

Measurement in the physical sciences. How can the place of measurement in any particular branch of science best be determined? Perhaps chief reliance must be placed upon the testimony of outstanding scientists in the particular field and recognized historians of science. Astronomy is doubtless the oldest and among the most highly developed of the sciences. Although the rise of experimental science is usually dated from Galileo, who lived about 300 years ago, Boring describes two important experiments in astronomy which were made as far back as 2,200 years ago. In commenting upon these early experiments, Boring¹³ says

It is no mere accident that these first two important astronomical experiments made use of mathematics in the interest of measurement. Measurement provides a precision of differentiation and definition in observation that can be had in no other way, mathematics provides the necessary means of carrying measurements through a logical development to their consequences without loss of their precision.

¹⁰ Kirtley F. Mather "The Common Ground of Science and Politics," *Science*, 117 167-174 February 20 1933 Page 170

¹¹ Bertrand Russell *The Impact of Science on Society* page 21 New York Simon and Schuster Inc., 1931

¹² Bertrand Russell *Dictionary of Mind, Matter, and Morals* page 227 New York Philosophical Library, 1932

¹³ Edwin G. Boring, *A History of Experimental Psychology* pages 14-15 New York The Century Company, 1929 Used by permission of Appleton Century Crofts Inc.

The bearing of this point upon the development of science is thus stated by Westaway ¹⁴

The more that exact measurement enters into any branch of Science, the more highly is that branch developed. It is for this reason that Chemistry and Physics are so far in advance of Botany and Geology. And the reason why we can obtain so much clearer notions of, for instance, an area or a weight, than of, say, wisdom or chivalry, is because the former are measurable, the latter not. It is of the first importance in Science that we should whenever possible, obtain precise quantitative statements of phenomena, and thus we see why it is that the introduction of a new scientific instrument so often leads to a marked advance in our knowledge.

Of the physical sciences, physics is usually regarded as the most highly developed at the present time ¹⁵. Two outstanding figures in the development of modern physics were Lord Kelvin in England and Max Planck in Germany. Regarding the place of mathematics and measurement in physics, Lord Kelvin says ¹⁶

When you can measure what you are speaking about, and can express it in numbers, you know something about it; and when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind. It may be the beginning of knowledge, but you have scarcely in your thought advanced to the stage of science.

Max Planck, regarded as one of the leading exponents of the important quantum theory in physics, declares that further progress in the physical sciences "will depend essentially on the development and wider application of our methods of measurement" ¹⁷. The case for measurement in physics may very well rest upon the testimony of these expert witnesses, although practically every prominent physicist from Galileo and Kepler to Einstein could be called.

Measurement in the biological sciences. The history of the various branches of science indicates clearly that not only are the biological sciences younger than the physical sciences, but also that measurement and mathematics have occupied and still occupy a less prominent place in the biological sciences. The primary reason for this is doubtless the greater simplicity of the data in the physical sciences ¹⁸.

¹⁴ F. W. Westaway, *Scientific Method: Its Philosophical Basis and Its Modes of Application*, pages 271-272. New York: Hillman Curl, Inc., 1937.

¹⁵ Silvio Fiala, "The Experiment and Its Role in the Theory of Knowledge," *Philosophy of Science*, 18, 203-208, July 1951. The highest level in the sense of most precise correspondence between its definitions achieved up to now is in physics, perhaps the oldest in natural science. Physics starts from the idea of the experiment as the only means by which this correspondence might be achieved. Page 251.

¹⁶ Quoted by Ronald King, "Physics, Metaphysics and Common Sense," *Scientific Monthly*, 42, 311, April, 1936.

¹⁷ Max Planck, *Where Is Science Going?*, page 96. New York: W. W. Norton & Company, Inc., 1932.

¹⁸ Julian Huxley makes this statement: "Sciences like Empires have their rise and their time of flourishing though not their decay. Naturally, the order of their rise runs parallel with the complexity of their subject-matter. The physical sciences, being the simplest and most straightforward, were the first to start their triumphant career." *What Dare I Think?* page 1. New York: Harper & Brothers, 1931.

Certainly, however, biology has progressed far beyond the stage of authority that existed in the Middle Ages, when, for example, the question of the number of teeth possessed by the horse was the subject of heated debate in many contentious writings "Apparently," says Loey, "none of the contestants thought of the simple expedient of counting them, but tried only to sustain their position by reference to authority"¹⁹ Loey recognizes three somewhat overlapping phases, or stages, in the development of the biological sciences namely, the descriptive, the comparative, and the experimental²⁰

Without doubt, one of the most important generalizations of modern science is that of evolution through natural selection as set forth by Charles Darwin near the middle of the last century, and yet it will be remembered that his work was nonmathematical, consisting largely of the classification of vast amounts of data upon which these epoch making generalizations were based In fact, as far as method goes, it was largely an extension and refinement of that emphasized by Aristotle more than two thousand years earlier

Whitehead attributes the retardation of science in the Middle Ages largely to Aristotle's emphasis on classification rather than measurement Note this statement²¹

But the biological sciences then and till our own time, have been overwhelmingly classificatory If only the schoolmen had measured instead of classifying how much they might have learnt!

That Darwin's cousin, Sir Francis Galton, took this position is clearly indicated by the following statement "Until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science"²² Galton, therefore, proceeded to introduce exact measurement and mathematical calculation into the theory of evolution In later years these biometrical methods have been greatly extended by Karl Pearson, Spearman, Fisher, and others The remarkable experiments of Mendel in heredity appeared at about the same time, although their value was not recognized until about 1900 Because of these pioneers, knowledge of heredity has become established on a definite mathematical basis²³ Even earlier than Mendel's time such noted physiologists as Müller, Weber, and Helmholtz had been doing much the same for physiology²⁴ After a survey of the development of natural science from

¹⁹ William A. Loey, *Biology and Its Masters* (Third Edition Revised), page 18 New York Henry Holt & Company, 1922

²⁰ *Ibid.* page 443

²¹ Alfred North Whitehead *op cit.* page 43

²² Sir Francis Galton quoted by I. W. Howerth on page 1 of *Measurement of Mental Phenomena* *The Delta Kappan* 15 1-9 June 2 1932

²³ Leslie C. Dunn (Editor), *Genetics in the 20th Century Essays on the Progress of Genetics during the First 50 Years* 634 pages New York The Macmillan Company 1931

²⁴ William A. Loey *op cit.* page 192

Aristotle to Fabre, which shows a definite trend from qualitative to quantitative analysis, Peattie concludes 'In short, what science calls for today are life histories, and ecological studies—the precise measurement of the environmental factors and the inter relations of organisms'²⁵

The various biological sciences have not been placed upon as definitely a quantitative basis as have physics and chemistry, largely because of the nature of their data. Some competent students of science think the biological sciences have moved too far in that direction. Whitehead²⁶ for example expresses regret that 'biology apes the manners of physics' while at the same time neglecting the unique character of its own subject matter—organisms, which are incapable of analysis without the destruction of their essential nature. The Gestalt school of psychology in recent years has also registered a vigorous protest against the atomic conceptions of mind which experimental psychology took over from nineteenth century physics.²⁷

Measurement in the social sciences Measurement in the social sciences presents a difficult problem. The social sciences are not only newer than the natural sciences but their data are more complex. They study human beings, the most complex of all biological organisms and their social relationships which are far more complicated than purely individual responses.

The genetic history of the social sciences has been described as follows²⁸

In the days of Aristotle Plato and Pythagoras philosophy still embraced the exact natural and social sciences. At the beginning of the nineteenth century the exact and natural sciences—mathematics astronomy physics chemistry geology, biology—had already left their philosophical matrix and were rapidly developing their own methods and techniques while preserving a tendency to return to philosophy for an occasional theoretical and speculative rehauling. But the social sciences—history ethics law economics psychology religion esthetics anthropology (such as it was)—were still rocking in the metaphysical cradle of Mother Philosophy. One by one the babes emerged and learned to stand on their own feet and to talk their own language even though their gait and vocabulary continued for a long time to bear traces of their maternal heritage.

In the preface to *The Seven Seals of Science*,²⁹ Mayer states his major thesis as follows

The central theme of the essay is that the sciences did not arise and could not have arisen simultaneously—that they form a well defined structure with mathematics at the bottom—that each later science built upon those that went before—that psychology is only now in process of becoming established and that the social

²⁵ Donald Culross Peattie *Green Laurels* page 345 New York Simon and Schuster, Inc. 1936

²⁶ Alfred North Whitehead *op cit* page 150

²⁷ David Katz *Gestalt Psychology Its Nature and Significance* (Translated by Robert Tyson) 175 pages New York Ronald Press Company 1930

²⁸ William F. Ogburn and Alexander Goldenweiser *The Social Sciences and Their Interrelations* pages 2-3 Boston Houghton Mifflin Company 1927

²⁹ Joseph R. Mayer *The Seven Seals of Science* page vii New York The Century Company 1927 Used by permission of D. Appleton Century Company

studies if they are to be worthy of the name of science, must build upon the natural sciences and particularly upon geology, biology, and psychology

It is significant that the author, although a professor of economics and sociology, uses as title for the final chapter in the book, "Social Science in the Making"

Other writers take a somewhat more optimistic position, and many of them indicate specifically the direction the development of social science is taking and must take. For example, Ogburn and Goldenweiser²⁰ write as follows

Attention finally must be drawn to the increasing importance of statistical methods in the social sciences. The extent to which social thought and theory will pass from the sphere of opinion, conjecture, and contemplative analysis to that of fact knowledge and control will depend on their permeation by these scientific methods of measurement and statistics

Of course there is nothing particularly new about this viewpoint. As early as 1798 Malthus attempted to put economics on a definite mathematical basis when he announced his celebrated, although erroneous, proposition that "population increases in a geometrical ratio, food in an arithmetical ratio." A little later, Quetelet showed that the theory of probability could be applied to human problems such as insurance.²¹

Barnes traces the history of sociology and concludes "There is a general agreement that sociology can become a true science of society only in the degree to which it is able to appropriate and apply those exact methods of measurement and analysis which constitute the indispensable attributes of science in general."²² On the other hand, Ellwood, an eminent sociologist, takes a wholly different position. His point of view is clearly stated in these words:²³

It would seem to me that as we ascend in the scale of life the view that science is quantitative measurement of objective conditions becomes less and less applicable not only because measurement becomes more difficult but because the subjective element plays a larger part. Even if the subjective element is capable of certain measurements and even if it is true that whatever exists exists in some quantity or number nevertheless it is obvious that where subjective elements play a large part measurement becomes of less importance for accurate knowledge because it is confined to the superficial aspects of the total situation and fails to expose the nature of the process which is being investigated. This is especially true in the social sciences and in them measurement seems to me to play a role secondary to other scientific methods

²⁰ William F. Ogburn and Alexander G. Goldenweiser, *op cit*, pages 8-9.

²¹ Helen M. Walker, *Studies in the History of Statistical Method*, pages 79-82. Baltimore: The Williams & Wilkins Company, 1924.

²² Harry Elmer Barnes, "The Development of Sociology," *Scientific Monthly*, 30: 547, December, 1932.

²³ Charles A. Ellwood, "The Uses and Limitations of the Statistical Method in the Social Sciences," *Scientific Monthly*, 37: 353-357, October, 1933. Page 353.

It seems fairly clear, therefore, that measurement and statistical analysis of quantitative data do occupy a prominent place in the social studies although there is no general agreement as to just what this place is. There does appear, however, to be universal recognition that the problems are more difficult than those presented by the earlier sciences and that their solution must be based at least in part on these other sciences, notably psychology. Measurement in psychology will be considered at some length in later sections.

Measurement in the applied sciences It has already been pointed out that the distinction between pure and applied science is not always easy to draw. In the beginning science appears to have arisen in the service of certain basic human needs and desires.³⁴ Despite its humble origin, however, science soon ceased to be but a means to an end and became an end in itself. For about a century and a half following Galileo it became exclusively the pursuit of the learned, and hardly influenced the thoughts or habits of ordinary men at all. The emphasis had shifted from applied to pure science. Russell comments upon this fact as follows:³⁵

It is only during the last hundred and fifty years that science has become an important factor in determining the everyday life of everyday people. In that short time it has caused greater changes than had occurred since the days of the ancient Egyptians. One hundred and fifty years of science has proved more explosive than five thousand years of prescientific culture.

The cycle is now complete. Science, which arose from man's stern necessity for meeting the ordinary problems of life, has now returned to serve again his practical needs. And nowhere is measurement more in evidence than in its practical applications. A competent observer³⁶ asserts that "one can hardly think of a field of intellectual endeavor into which measurement has not crept, and surely there is none in which its influence has not been felt."

Indeed, it is these practical applications of science that have impressed the mind of the layman in recent years. When he hears the word "science" he is likely to think of the results of science, possibly some invention such as the radio or radar, rather than of the physics and chemistry that have made them possible. Probably a thousand persons would know of Marconi, who invented wireless telegraphy, to one who ever heard of Hertz or Maxwell, whose pioneer work blazed the trail.

The prominent place of quantitative measurement in these modern applications of science to engineering and industry is too well known to require elaboration. Take a modern automobile as an example. Its mechanical parts are accurate to the thousandth part of an inch. Every detail has been

³⁴ Cf. John Dewey, *How We Think*, page 216. Boston: D. C. Heath & Company, 1913.

³⁵ Bertrand Russell, *The Scientific Outlook*, *op cit*, pages vii-viii.

³⁶ B. Othmel Smith, *op cit*, page 2.

subjected both to careful laboratory experimentation and to rigid tests on the trial grounds. The instrument board presents, as practical aids to the user, various devices for measuring gasoline, electric current, oil pressure, temperature, and speed of car, as well as perhaps a clock and a radio.

It is instructive to study what the application of science has done for modern cookery. The ordinary untrained housewife still uses recipes with such vague directions as "season to taste," "add butter to the size of a walnut," "cook in a moderate oven," and so on. In contrast, the modern bakery accurately measures all ingredients, mixes them uniformly for a specified length of time, and then cooks them at a specified temperature for a definite time. This assures a predictable uniformity in the product, in contrast with the "luck" of the old time cook.

Medicine is an outstanding field in which many discoveries of pure science have been applied to the solution of practical human problems. Herrick³⁷ explains how the development and use of various instruments of precision have revolutionized medical diagnosis and practice. The measurement of blood pressure, body metabolism, and the physical and chemical analyses of the blood and other body fluids are as recognized techniques today as were height, weight, and temperature a generation ago. The dietitian in the kitchen measures the patient's food from the standpoint of calories, minerals and vitamin content with an accuracy approaching that of the pharmacist in compounding his medicines and of the nurse in administering them.

Limitations of measurement. Before concluding this discussion of the relation between measurement and science, it may be well to note some of the difficulties and problems of measurement. It must not be assumed that the tools and techniques of measurement have been developed to a state of perfection. This is far from true even in physics and chemistry, where measurement has progressed furthest.

Planck, an eminent physicist, offers the warning that "every number obtained by physical measurements is liable to a certain possible error."³⁸ Westaway puts the matter in these words: "We may, in fact, look upon the existence of error in all measurements as the normal state of things."³⁹ Doubtless, Bertrand Russell has the same idea in mind when he describes science as a "succession of approximations."⁴⁰

In general it may be said that the sources of error in measurement are due to the imperfections either in the measuring instruments themselves or in the method in which they are employed. While both of these sources of error are subject to a considerable measure of control, neither can be

³⁷ James B. Herrick, *Changes in Internal Medicine Since 1900*, *Journal of American Medical Association* 100: 1312-1315, October 26, 1935.

³⁸ Max Planck, *A Survey of Physics*, page 92. New York: E. P. Dutton & Co., Inc. 1925.

³⁹ F. W. Westaway, *op cit*, pages 289-290.

⁴⁰ Bertrand Russell, *The Scientific Method*, *op cit*, page 63.

eliminated altogether Three methods of controlling errors in measurement may be suggested:

- 1 The improvement of existing measuring instruments
- 2 The devising of adequate methods of *estimating or allowing for errors*
- 3 The development of skill in applying the instruments of measurement so as to reduce errors to a minimum and in interpreting the results so as to take due account of the errors which cannot be eliminated

The first of these methods will be considered at some length in Chapters 4 to 7, the second will receive attention in Chapter 3, while practically the entire book is concerned with the third

The limitations of existing measuring instruments do not detract from the importance of measurement although they do add to its difficulty The result is rather to set a special premium upon the skillful use of these instruments As a rule, the cruder any tool is the greater the skill required in its application, if satisfactory results are to be obtained The early automobile, for example, called for much greater skill in its successful operation than does its more highly perfected modern successor

Conclusions. What, then, is the relation between measurement and science? A few generalizations seem fairly clear

1 There is a direct relationship between the status of a science and the degree to which measurement has been developed in it In the older and better established physical sciences, measurement occupies a fundamental place, in the newer biological sciences, measurement occupies a less important place, and in the social sciences, the most recent group measurement has made hardly more than a beginning The evidence seems abundantly to support Westaway's statement "The more that exact measurement enters into any branch of Science, the more highly is that branch developed"

2 The prominence of measurement in a science appears to be roughly in inverse ratio to the complexity of its subject matter Inert material seems inherently more susceptible to measurement than living organisms Apparently the maximum difficulty comes in the case of man particularly in his social behavior

3 All measurement is subject to errors These errors are due to limitations in the tools as well as in the techniques of measurement To ensure satisfactory results the greater the limitations of the former the greater the skill and insight called for in the latter

B. Measurement in Education

Is education a science? Education is described sometimes as a philosophy, sometimes as a science, and sometimes as an art Moreover, it is

^a F W Westaway, *op cit* pages 271-272

commonly assumed that these terms can be clearly distinguished from each other or even that there is a certain antagonism among them. Quite contrary is the truth, however. They are most closely interrelated. As the role of measurement in education will doubtless be different if education is a science from what it will be if education is a philosophy or an art, some attention must now be given to the problem of the relationship of science to philosophy on the one hand and to art on the other. Benjamin's comments are pertinent here:

It is commonly recognized that the most significant difference between philosophy, on the one hand and the more specialized studies, such as science, art, religion, education, and politics, on the other, is that philosophy is concerned with the critical examination of certain concepts and beliefs that are presupposed by these more specialized studies. For example, the average scientist readily admits that although his study is based upon certain beliefs in the uniformity of nature, the rationality of the world, and the relative attainability of truth, he is not himself, as a scientist, called upon to subject these notions to critical examination. He can argue either that such justification is not ordinarily required for the actual carrying on of science, or that if such justification is demanded it can readily be provided by the philosopher whose job is, by common consent, precisely the examination of the assumed notions of science.

Perhaps the kinship of science and philosophy will be more apparent if approached genetically and historically. The early Greek thinkers, Plato and Aristotle for example, recognized no distinction between science and philosophy, both being joined by the common bond of love, the pure love of truth. During the Middle Ages, however, this once harmonious family found itself unequally and somewhat unhappily yoked together. Since the later Renaissance, one after another of the children born to this union, beginning with physics and chemistry, left the family tree and set up in business for themselves. This was a sort of "psychological weaning," which doubtless performed a useful function for the time being. But, as often happens when discontented youth desires to assert its independence, science rebelled altogether and would have nothing at all to do with the parental wisdom of Mother Philosophy. This prolonged period of arrogant adolescence has continued to the present time. Fortunately, in recent years some of the wiser heads have somewhat patched up the family quarrel which has brought unhappiness to mother and daughters alike. Gray describes the result in education:

- "A. Cornelius Benjamin. Science and Its Presuppositions." *Scientific Monthly* 73: 150-153 September 1931. Page 150.
- "Harold R. Smart. *op cit* Chapter I.
- "For an excellent general discussion of this point of view see Max Black. 'A Lend Lease Program for Philosophy and Science.' *Scientific Monthly* 61: 165-172 September 1915.
- "For an admirable statement of the dangers of a narrow conception of a science of education see William A. Brownell. 'Quantitative Research on Teaching and Learning.' *School and Society* 50: 847-856 December 30 1939.
- "J. Stanley Gray. 'A Neglected Phase of Educational Research.' *Journal of Educational Research* 29: 89-90 October, 1935.

Educators are so over-zealous to become "scientific" and objective that they have ignored the fact that there is no less need for thinking in science than in philosophy. When research procedure neglects theoretical evaluation and interpretation, it is only partly scientific.

Although competent students of education recognize that both philosophy and science are necessary for a complete act of thinking, each field is so broad as to make a certain division of labor necessary. The situation has been well stated as follows:⁴⁶

If one specializes in the critical examination of educational theories, hypotheses and generalizations in the light of data which are already available we call him an educational philosopher. If one specializes in the solving of educational problems by making new appeals to experience through systematic, controlled and uncontrolled observation, in field or laboratory, we call him an educational scientist in the classical sense of the term.

But even the most competent philosopher is forced to take his science at second hand from the data made available by specialists in science, for it is too much to expect him at the same time to be an expert experimentalist.⁴⁷ Conversely, a competent scientist is forced to take his philosophy largely at second hand while he conducts his persistent search for new facts. Moreover, not only does a scientist have to borrow his philosophy, but much of his science also.

This borrowing, though necessary, is risky business not only for the philosopher but for the scientist as well. Not only may he be unfortunate in what he borrows, but its meaning to him is inevitably colored by the mental background imposed by his specialty. The important point is that science and philosophy are reciprocally related as inseparably linked as are heredity and environment in the growth of a living organism. Buckingham states the relationship concisely: "As fields of human endeavor science and philosophy supplement each other. Without philosophy, science is incomplete, without science, philosophy is barren."⁴⁸ An educational philosopher puts the relationship as follows: "While philosophy must be the general to plan the grand strategy of education, it will need science as its staff officer."⁴⁹ Broadly conceived, education then is both science and philosophy. As a science it belongs to the group known as the social sciences, whose data are the most complex of all. While education is not, and doubtless never will be, as thoroughgoing a science as physics or chemistry, it has nevertheless made more progress toward the scientific treatment of its

⁴⁶ Carter V. Good, A. S. Barr and Douglas E. Seates, *The Methodology of Educational Research*, page 24. New York: D. Appleton-Century Company, 1936.

⁴⁷ In commenting upon William James Boring makes this disturbing observation: "It is too bad, but no one has ever yet succeeded in being both a good philosopher and a good experimentalist." E. G. Boring, *op. cit.*, page 502.

⁴⁸ B. R. Buckingham, "The Philosophy and Organization of Research," *School and Society*, 29: 758, June 15, 1929.

⁴⁹ John S. Brubacher, *Modern Philosophies of Education*, page 67. New York: McGraw-Hill Book Company, Inc., 1939.

problems during the twentieth century than in all the centuries preceding.

But what of art in relation to science and philosophy? Will Durant puts the matter clearly and graphically.⁵⁰

Every science begins as philosophy and ends as art, it arises in hypotheses and flows into achievement. Philosophy is the front trench in the siege of truth. Science is the captured territory, and behind it are those secure regions in which knowledge and art build our imperfect and marvelous world.

William H. Payne long ago suggested that "in the slow but sure evolution of human opinion, a science of education is beginning to emerge from the art of education."⁵¹ But there is also a reciprocal relationship, for as Doughton⁵² points out, "the sure foundation of an effective teaching art is a science of education."

Science consists in *knowing*, while art refers to *doing* and implies skill and aesthetic excellence. An outstanding teacher might be either an artist or a scientist, although the ideal teacher must be something of both. No matter how great the artist or with how much inspiration he wields the brush, the pigments are mixed according to formula.

The place of measurement in education. What, then, is the rightful place of measurement in education, which is at once a science, a philosophy, and an art? The answer varies somewhat with the point of view of the observer. Naturally the role of measurement appears more important to those educators whose specialty is science than to those whose specialty is philosophy. At times these views become so divergent as to appear wholly irreconcilable. The quotations at the top of the next page, from outstanding early educational leaders in a single institution, will make this clear.

It is always dangerous, of course, to detach a statement from its context. However, the strong language employed, containing such terms as "thoroughly," "indispensable," "final," "fallacy," "never," and "always," scarcely leaves room to hope that these statements are entirely harmonious. Doubtless, any attempt to interpret the above quotations should take into consideration the publication dates. It will be noted that views of the educational scientists represent an earlier period. They appeared soon after World War I, when the success of the Army Alpha test was still fresh in the minds of psychologists. Moreover, at that time, atomic physics still dominated all science, including psychology, and behaviorism was in the ascendancy in America. The relatively more recent quotations from educational philosophers reflect a newer atmosphere. The recognition of the principle of indeterminacy in physics has had a sobering effect on science in general, and followers of the Gestalt school of psychology in particular.

⁵⁰ Will Durant, *The Story of Philosophy*, pages 2-3. New York: Simon and Schuster, Inc., 1926.

⁵¹ William H. Payne, *Contributions to the Science of Education*, page 11. New York: Harper & Brothers, 1886.

⁵² Isaac Doughton, *Modern Public Education: Its Philosophy and Background*, page 185. New York: D. Appleton-Century Company, 1935.

TWO VIEWS OF MEASUREMENT IN EDUCATION

EDUCATIONAL SCIENTIST	EDUCATIONAL PHILOSOPHER
Whatever exists at all exists in some amount To know it thoroughly involves knowing its quantity as well as its quality ⁵³	Yet another false concept in the climate gripping the American scholar thwarts his study of Man This is the fallacy that "I know only what I can describe quantitatively whatever exists, exists in some measurable amount" ⁵⁴
Measurement is indispensable to the growth of scientific education The final answer to every educational question, except one, must be left to the educational measurer and must await the development of education as a science ⁵⁵	And I should myself like further to conclude that education can never become a science always—so long as this world stands—will there be problems, nay regions of problems with which the processes of "exact" science are insufficient to cope ⁵⁶

have protested against what they regard as the atomic conception of mind It is, therefore, quite possible that the present views of its two groups are much closer together than the above statements would indicate However, the following statement from McCall strongly suggests that not all differences have been ironed out⁵⁷

Certain extreme exponents of the organismic (often called *Gestalt*) view contend that any organism is more than the sum of its parts, and that adding test scores is like trying to make a man by sticking together a head a trunk, two arms and two legs But a reading score cannot be properly compared to one leg It is not a broken off fragment of the mind In a very real sense, a reading score tends to measure the entire organism functioning in that reading situation

Mental measurements are essentially similar to bodily measurements If anyone proposed to abolish the making and use of the measurements of pulse temperature blood pressure, et cetera we would call him *crazy*, and if any one proposes to abolish the making and use of mental measurements, he, too, should be called—I hesitate to say what, since somehow I must manage to live with certain of my colleagues after this is published, but surely something other than an organismic philosopher or a *Gestalt* psychologist!

Both scientists and philosophers attempt to test their generalizations before finally accepting them With science the process is the straightforward one of subjecting all such generalizations to rigid mathematical or

⁵³ Edward L Thorndike *Seventeenth Yearbook of the National Society for the Study of Education Part II*, page 16 1918

⁵⁴ Harold Rugg "The American Scholar Faces a Social Crisis," *The Social Frontier* 1 12 March 1935

⁵⁵ William A McCall *How to Measure in Education* pages 7, 9 New York The Macmillan Company, 1923

⁵⁶ William H Kilpatrick *School and Society* 30 48 July 13 1929

⁵⁷ *The Test Newsletter* published by the Bureau of Publications Teachers Coll ge Columbia University, December, 1936

experimental verification. With philosophy the process appears more involved. Kilpatrick,³³ for example, recognizes two distinct situations "simple prophecies" and "decisions on appropriate conduct or policy." For the former, he agrees that "verification" is an appropriate term and measurement (when available) is a proper means of testing. For the latter, however, he insists that "verification" is not an appropriate term and techniques of measurement are not in themselves adequate. He continues "In such cases the function of measurement is not to supplant or to supply decisions, but to furnish regarding the working of the policy under review, more and better data in the light of which a fresh and better decision can be made." Apparently, then, whenever actual verification is possible, this philosopher at any rate is willing to assign to measurement the job of doing it, and even in the other cases he assigns to it the necessary, if humble, duty of providing at least part of the data required. Perhaps, then, it is not an unfair statement to say that the scientist *always* assigns to measurement a fundamental role, whereas the philosopher *sometimes* does so. However, at all times the philosopher seems willing to ascribe to measurement an important, even if not a fundamental, place in education.

Take the important matter of guidance, for example. Even its most enthusiastic supporter would hardly characterize guidance as a full fledged science. Yet measurement provides some of the essential data in any sound guidance program. To describe a pupil as of weak scholarship and of low mentality is to leave his status vague and unsatisfactory. But to say that he has a percentile rank of 20 on the California Achievement Tests and an IQ of 84 on the Revised Stanford Binet Intelligence Scale is to describe him in reasonably precise and meaningful terms.

In this connection it is well to observe that measurement is always a means to an end and never an end in itself. A measurement is simply a quantitative description of observed data. The significance or educational implications of the measurement are rarely self-evident or automatic. As a rule, the true significance of the measurement can be determined only when it is seen in relation to other relevant factors and is fitted into the total pattern of the situation. The term *evaluation*, as distinguished from *measurement*, is often used to refer to the process of appraising the "whole" child or the "entire" educational situation.

The three R's in education. Everyone is familiar with the famous trinity of R's in education, "Readin', 'Ritin', and 'Rithmetic." These, of course, have to do with the content of education, the curriculum of instruction. There is also another series of R's which is concerned with the process of arriving at, or at least of searching for, truth in education to serve as a basis for theory and practice. Educators have, in general, employed three principal methods of settling educational issues and of arriving

³³ William H. Kilpatrick, "The Relation of Philosophy to Scientific Research," *Journal of Educational Research* 24: 110-111, September 1931.

at educational principles and policies. These constitute a new series of R's, "Rhetoric, Reputation, and Research."

Historically the first of these methods may be termed that of Rhetoric. It is the method *par excellence* of politicians, although not unknown in education, especially among the reformers of every period. The method is usually most dangerous when used orally. It is too well known to require detailed discussion here. Abe Martin's famous definition of an orator as a "public speaker not unduly hampered by the facts" indicates rhetoric's limitations. The danger is that the personality of the speaker may outweigh the merits of the case, and the artistic form of the speech may have more influence than its content.⁵⁹ Naturally measurement and quantitative data are usually irrelevant, if not positively in the way. As a matter of fact, a Speaker of the House of Representatives once attributed the decline in oratory chiefly to the "general diffusion in knowledge," since 'as a rule the more information a man has the less emotional he is, and the orator's appeal was to the emotions far more than to the understanding.'⁶⁰

The second method of determining educational theory and practice may be termed that of Reputation. According to it, the settlement of an educational issue is the simple matter of finding out what has been said on the question by some persons whose reputations in the field are sufficiently great to make them accepted as authorities. This method has been the dominant one in education until recently and is still widely used. It has a legitimate and necessary place in education as it does in law and medicine where one must rely for the solution of many practical problems upon the professional judgment of acceptable authorities. But the method is not without its dangers, which are so important as to warrant brief discussion.

In the first place, the authority may be mistaken. Reputation is unfortunately no guarantee of reliability. Until comparatively modern times the wisest persons were quite certain that the sun each day made a complete journey around our flat earth. Such divergent views on practically every phase of education as are expressed in current educational journals and in public addresses by our most eminent educational leaders are ample assurance that men of the highest reputation may be mistaken.⁶¹ In the second place, the authority may be misquoted. A few years ago at a meeting of the American Psychological Association a speaker quoted what pur-

⁵⁹ Irvin S. Cobb described a prominent Southern orator as one who can 'make a song of a syllable and turn any reasonably long word into an anthem.' *The Courier Journal* Louisville Kentucky June 16 1936.

⁶⁰ Chauncy Clark, 'Is Congressional Oratory a Lost Art?' *Century* 81 310 December 1910.

⁶¹ Educators of course have no monopoly here. Bertrand Russell reports the amusing but instructive example of Todhunter the mathematician who opposed the establishing of the first experimental laboratory at Cambridge because he thought it was unnecessary for students to see experiments performed since the results could be vouched for by their teachers all of whom were men of the highest character including many who were clergymen in the Church of England!

ported to be a statement from an outstanding psychologist. At the conclusion of the address, this psychologist arose to explain that he had never made any such statement and in fact believed quite the contrary. It is too much to expect, however, that the authority will always be present to correct his alleged quotations. A third danger is that conditions may have changed so greatly that a statement once true may no longer be applicable.⁴⁰ For example, George Washington's warning against foreign entanglements, made in 1797, when the United States consisted of 16 states whose total population, barely 5,000,000, was separated from Europe by a broad Atlantic, need not be at all applicable to a nation of 48 states, with a population of more than 150,000,000 joined to Europe by modern agencies of communication. The method is thus seen to be beset by many dangers. It must, therefore, be used with caution. The necessity of extreme care in the selection of the authority cannot be overemphasized. It is usually wise, also, to examine the evidence that lies behind the statement and the conditions under which it was made. Reputation alone must not be thought of as adequate assurance of reliability. At best the reliance upon reputation involves considerable risk. It is always well to consider the circumstances under which the statement was made as well as the data upon which it was based.

The third method of arriving at truth is that of Research. This method is comparatively recent in the history of man. The prestige of the orator and rhetorician in ancient Greece and Rome and the authority of Aristotle in the Middle Ages testify to the newness of the method of research. And it is more recent in education and the other social sciences than in the physical and biological sciences. Its appeal is to the intellect and is based upon the facts in the case. It is the distinctive method of science and may be regarded as the only final method of settling an educational issue. A single illustration will indicate its superiority over the earlier methods used.

A practical problem in education is to determine the proper amount of time to be devoted to each subject taught. Educators have usually assumed that the results obtained are directly proportional to the amount of time expended. In fact, the thing seemed self-evident. In the closing years of the last century, however, an inquisitive American physician by the name of Rice undertook, apparently for the first time, to subject the question to scientific study. The subject chosen was spelling, and the procedure was extremely simple and direct. A uniform, although not standardized, spelling test was administered to schools in various parts of the country. Afterward the results, involving about 100,000 cases, were tabulated according to the amount of time devoted to spelling in the school program. Contrary to the usual assumption, Rice found little or no relation between the results

⁴⁰ Someone has suggested that theories often continue to live long after their brains have been knocked out.

obtained and the time expended.⁶³ Equally good spelling achievement was found in schools where a period of ten or fifteen minutes was devoted to the subject as in those where a period three or four times as long was allowed. Although considerable skepticism was manifested toward the Rice inquiry at the beginning, the evidence was so convincing as to compel assent. Today few schools allot more than fifteen minutes a day to spelling in the school program. The solution of practical problems in education by the method of research had thus made a promising beginning.

Forty one years later Tyler summarized the educational situation as follows.⁶⁴

The proceedings of educational associations during the latter part of the nineteenth century indicate clearly an attempt to settle teaching problems by argument by impassioned pleas, or by consensus. The achievement-testing movement provided a new tool by which educational problems could be studied systematically in terms of more objective evidence regarding the effects produced in pupils. The hope that problems could be settled by reference to fact rather than subjective impression or emotionally colored opinions has probably been the strongest influence of the achievement-testing movement in the past forty one years.

Good, Barr, and Scates⁶⁵ classify research methods in education under four headings, historical, normative-survey, experimental, and other methods. Of these four methods only the historical is not dependent upon measurement in some form, and even this method is likely to make use of numerical data.

The function of measurement in instruction and in school administration. The foregoing discussion has been concerned primarily with the role of measurement in educational research, or in education considered as a science. But education is an art as well as a science. It has its practical as well as its theoretical aspects. It is a primary purpose of this book to consider such immediately practical problems as the relationship of measurement to the actual administration of schools and the instruction of pupils in these schools. Later chapters will consider measurement in in-

⁶³ Joseph M. Rice, 'The Futility of the Spelling Grind,' *Forum* 23: 163-172, 403-419, April and June 1897. Later studies have found a similar lack of relationship in other subjects. See Merrill T. Laton, 'A Survey of the Achievement in Social Studies of 10 220 Sixth Grade Pupils in 464 Schools in Indiana,' *Bulletin of the School of Education, Indiana University* 20: No. 3, 1944, 68 pages.

Rice was an important progressive education pioneer. During the years 1891-1899 he investigated American education relentlessly and published a series of 20 articles in the *Forum* which even today sound startlingly modern. The first was 'Need School Be a Blight to Child Life?' (12: 529-535, December 1891), the last 'Why Teachers Have No Professional Standing' (27: 452-463, May 1899). For further historical perspective see Douglas C. Scates, 'Fifty Years of Objective Measurement and Research in Education,' *Journal of Educational Research* 41: 211-234, December 1947.

⁶⁴ *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, 'The Scientific Movement in Education'*, page 349. Quoted by permission of the Society, Bloomington, Illinois: Public School Publishing Company, 1938.

⁶⁵ Carter V. Good, A. S. Barr, and Douglas L. Scates, *op. cit.* 82 pages.

struction and measurement in school administration. There is some overlapping among the two major divisions, each of which will be subdivided still further. But the organization is a convenient one, even if somewhat arbitrary.

It must be admitted at the outset that up to the present time research, while not entirely lacking, is by no means sufficient to prove *conclusively* that measurement really serves the above practical functions. However, though experimental evidence for the practical value of measurement is somewhat meager, objective support for the view that measurement is useless or harmful is virtually nonexistent. The present case for measurement in education rests to a large extent upon the testimony of experienced teachers and school administrators, and the argumentative ability of persons enjoying the highest reputation in the field.⁶⁵ This is one of the many points in education where further research is needed.

Examples of existing experimental evidence as to the value of measurement in instruction are the following:⁶⁶

Schutte found that normal school students who expected final examinations did significantly better than those who did not. Kulp found weekly tests increased the amount learned in educational sociology by about 17 per cent. Turner found that educational psychology students who took twelve short tests did about 20 per cent better than others who took only the mid term and final examinations. Jones found that psychology students who took a five-minute test after each lecture retained after eight weeks approximately twice as much as those who did not. Keys found that the same tests administered in the form of weekly rather than monthly examinations in educational psychology gave an immediate superiority of 12 per cent.

Great strides have recently been taken at all school levels—elementary, secondary, and college—toward relating tests to educational objectives. These are set forth in considerable detail by 20 specialists in a highly important book, *Educational Measurement*.⁶⁷ Trends in the research literature are summarized frequently by the *Review of Educational Research*, whose issues contain rather comprehensive bibliographies.⁶⁸

⁶⁵ Three typical articles illustrating this point of view are: Hans C. Gordon, "How Teachers Use Test Scores to Understand the Needs of Pupils," in *Growing Points in Educational Research*, 1949 Official Report of the American Educational Research Association, pages 276-278, Washington, D. C., National Education Association, 1949; Aileen Lockhart, "Testing Can Improve Teaching," *Journal of Health and Physical Education*, 19, 627-629, November, 1948; Jane E. McAllister, "Tests and Measurements Aid Human Relations," *Educational Administration and Supervision*, 30, 49-58, January, 1949.

⁶⁶ These studies and other related studies will be discussed more fully in Chapter 11.

⁶⁷ L. F. Lindquist (Editor), *Educational Measurement*, Washington, D. C., American Council on Education, 1951, 819 pages.

⁶⁸ Paul Blommers (Chairman), "Methods of Research and Appraisal in Education," 21, 327-401, December, 1951; J. Raymond Gerberich (Chairman), "Educational and Psychological Testing," 20, 1-99, February, 1950; Frederick B. Davis (Chairman), "Educational and Psychological Testing," 23, 1-110, February, 1953.

Types of measurement in education The various types of measurement employed in education may be classified on different bases and from different points of view. The following appears to be a reasonably satisfactory classification of the instruments of measurement employed in the ordinary school for distinctly educational purposes

A Oral

B Written

1 Informal (nonstandardized)

a Essay

b Objective

2 Formal (standardized)

a Achievement

(1) General (survey)

(2) Specific (diagnostic practice etc.)

b Intelligence

(1) General (individual and group)

(2) Specific (aptitude or prognosis)

c Personality and Interests

The distinction between the major categories oral and written is obvious. The distinction between informal and formal written tests is also easy to make. A formal test often begins as an informal test which is later subjected to experimental trial and revision only the best items surviving the process. Formal tests also have carefully worded instructions both for administering and scoring and usually norms for interpreting the results.

The distinctions among tests of achievement intelligence personality and interests are not so clear cut however. By the term *achievement tests* is meant tests of academic achievement such as arithmetic or algebra they are distinguished from most personality and interest tests which are self report rather than ability measures.⁷⁰ Intelligence tests theoretically at least, are measures of learning capacity whereas achievement tests are measures of learning itself.⁷¹ In other words intelligence tests attempt to measure educability while achievement tests attempt to measure education. The writers have followed the usual practice of recognizing tests of achievement and intelligence as co-ordinate with tests of personality and interests. Strictly speaking however achievement and intelligence are merely aspects of personality which is a term used by psychologists to include every trait that differentiates one individual from another. In a certain sense then every test is a test of personality and many aspects of personality cannot be measured by tests at all but are evaluated by means

⁷⁰ Lee J. Cronbach *Essentials of Psychological Testing* pages 14-15 New York Harper & Brothers 1949

⁷¹ Tilton comes to the interesting conclusion that his data are fairly consistent with a concept of a general ability to learn and with the identification of it with the general intelligence test. John W. Tilton The Intercorrelations between Measures of School Learning *Journal of Psychology* 30 169 173 January 1933 Page 173

of rating scales, questionnaires, interviews, controlled observation, and the like

Tests are subdivided into general and specific on the basis of scope. They may also be further subdivided into individual and group tests on the basis of method of administration and into verbal and nonverbal or performance tests on the basis of content. A distinction is often made, although not always observed in practice, between a test and a scale. A test consists of a series of questions to be answered or exercises of some sort to be done, and a pupil's performance is the number of these he is able to do in the time allotted. Strictly speaking a scale consists of a series of specimens such as handwriting for example arranged in order of merit, and the pupil's performance is judged by comparing it with the standard specimens. Most standardized instruments of measurement are really tests rather than scales. The test items are frequently arranged in order of difficulty, however in which case the term *scaled test* is sometimes used to distinguish such tests from those in which the items are not so arranged.⁷²

Of course, many other types of measurement are employed in schools. Examples of these are chronological age, height, weight, temperature, and time but these can hardly be classified as strictly educational. It cannot be too strongly emphasized that measurement is not limited to tests and examinations and certainly not to standardized tests. There are also numerous rating scales and check lists for playgrounds, buildings and equipment and so forth, whose use is largely restricted to the specialist. These have been omitted in the interest of brevity.

It must be recognized that recent tendencies in education have enlarged its scope and increased its complexity, and have thereby added to the difficulties of teaching and administration. But nowhere have these difficulties been more apparent than in the problem of measurement. The need for proper evaluation is as great in the modern school as ever before but the difficulties of providing for it are vastly greater. For, as Saucier points out, an instrument of measurement may meet all the criteria for measuring a reactionary undemocratic conception of education but at the same time be valueless for measuring the major results of a progressive, democratic theory of education.⁷³ This means that as the schools improve, so must the tools and techniques of measurement and evaluation.

A quotation from Scates seems to sum up the central theme of this chapter quite aptly.⁷⁴

⁷² For an extensive discussion of modern scale analysis see Samuel A. Stouffer and others, *Measurement and Prediction*. Volume IV of *Studies in Social Psychology in World War II*. 700 pages. Princeton, N. J. Princeton University Press, 1950.

⁷³ W. A. Saucier, *Introduction to Modern Views of Education*, pages 368-369. Boston: Ginn & Company, 1937.

⁷⁴ Douglas E. Scates, "Fifty Years of Objective Measurement and Research in Education," *Journal of Educational Research* 41: 241-264, December 1947. Pages 253-254.

What has the measurement movement done for us? One way of responding to this question is to note what we should not have if we had no such measurements. Certainly we would not have much of our current scientific work in education. For there cannot be a science without fairly precise quantification, not that science is measurement, but that traits which are devoid of any reasonably definite quality simply do not have the required specificity for entering into the careful thinking essential to science. When quantities are disregarded almost any generalization becomes true. There is an infinitesimal element of truth in practically anything which might be said. Quantities are part of the nature of truth and are therefore an essential part of science.

SELECTED REFERENCES FOR FURTHER READING

(References cited in the footnotes to Chapter 1 are not repeated here)

- Barr, Arvil S., Davis, Robert A., and Johnson, Falmer O., *Educational Research and Appraisal* New York J B Lippincott Company, 1953 362 pages
- Boring, Edwin G., *A History of Experimental Psychology* (Second Edition) New York Appleton-Century-Crofts, Inc., 1950 Chapter I, "The Rise of Modern Science," and Chapter XXIII, "Gestalt Psychology"
- Brubacher, John S. (Editor), *Eclectic Philosophy of Education* New York Prentice Hall, Inc., 1951 Chapter II, "Science and Philosophy of Education Compared"
- Cohen, I. Bernard, *Science Servant of Man, a Layman's Primer for the Age of Science* Boston Little, Brown and Company, 1948 Part I, "The Nature of the Scientific Enterprise"
- Davis, Frederick B., "The AAF Qualifying Examination" *Army Air Forces Aviation Psychology Program Research Report No 6* Washington, D C U S Government Printing Office, 1947 266 pages
- Dewey, John, *Logic, The Theory of Inquiry* New York Henry Holt & Company, 1938 Chapter XI, "The Function of Propositions of Quantity in Judgment" and Part IV, "The Logic of Scientific Method"
- Einstein, Albert, *Out of My Later Years* New York Philosophical Library 1950 282 pages
- Gulliksen, Harold, *Theory of Mental Tests* New York John Wiley & Sons, Inc., 1950 486 pages
- Jahoda, Marie, Deutsch Morton, and Cook, Stuart W. (Editors), *Research Methods in Social Relations, with Especial Reference to Prejudice* New York The Dryden Press, 1951 Part I, "Basic Processes," and Part II, "Selected Techniques"
- Kemble, Edwin C., "Reality, Measurement, and the State of the System in Quantum Mechanics," *Philosophy of Science*, 18 273-299, October, 1951 "At the heart of modern experimental science is the controlled experiment in which objectivity is secured by reducing all measurement to the reading of scales of one sort and another. Good experimentation, however, requires more than the replacement of qualitative subjective observations by pointer readings. It requires careful analysis of all possible perturbing factors and repeated measurements to test the scatter of the results" (page 274)
- George, Irving "The Fundamental Nature of Measurement," Chapter XIV in E F Lindquist (Editor) *Educational Measurement* Washington, D C American Council on Education, 1951

- Planck, Max, *Scientific Autobiography, and Other Papers* New York Philosophical Library, 1949 192 pages
- Ruby, Lionel, *Logic, an Introduction* New York J B Lippincott Company, 1950 Part III, 'The Logic of Truth Scientific Methodology "
- Russell, Bertrand, *A History of Western Philosophy* New York Simon & Schuster, 1945 895 pages
- Sarton, George, *The Life of Science Essays in the History of Civilization* New York Henry Schuman, 1948 Part I, "The Spread of Understanding "
- Stevens, S Smith, "Mathematics, Measurement, and Psychophysics," Chapter I in S Smith Stevens (Editor), *Handbook of Experimental Psychology* New York John Wiley & Sons, 1951
- Whitney, Frederick L, *The Elements of Research* (Third Edition) New York Prentice-Hall, Inc, 1950 Chapter I, 'Reflective Thinking, Science, and Research "
- Williams, Donald, *The Ground of Induction* Cambridge, Mass Harvard University Press, 1947 Chapter V, "The Logic of Science "

2

The Historical Development of Measurement in Education

A Introduction

Tests and measurements of one kind or another have played a far more prominent role in human history than is generally recognized. Nor has their use by any means been confined to the schools. In fact among the earliest records of the use of various testing devices are those found in the Bible, although they generally have no direct reference to education. One illustration¹ will suffice.

And the Gileadites took the passages of Jordan before the Ephraimites and it was so that when those Ephraimites which were escaped said Let me go over that the men of Gilead said unto him Art thou an Ephraimite? If he said Nay then said they unto him Say now Shibboleth and he said Sibboleth for he could not frame to pronounce it right Then they took him and slew him at the passage of Jordan and there fell at that time of the Ephraimites forty and two thousand.

Attention is called to the fact that here is indeed a "final examination and in a field other than education. Doubtless measurement experts of the present time would point out that, in spite of a rather high degree of objectivity, there were certain dubious features: it was oral, it was very short and the mortality rate was excessively high."

A sociologist² attributes the remarkable stability of the Chinese civilization, the oldest culture of any modern nation to five factors one of which is her highly organized examination system. It began informally in 225 B.C.

¹ *Judges 12 5-6 (King James Version)*

² Paul F. Cressey, "The Influence of the Literary Examination System on the Development of Chinese Civilization," *American Journal of Sociology* 30: 250-262, September 1929.

and became a definite civil service examination system in 29 B.C. The system, described as being thoroughly democratic, ruthless, invariable, and orthodox, has had profound effects, some good and some bad, not only upon the educational system of China, but also upon her whole civilization. On the one hand, it has preserved unity by keeping uniform throughout the empire the written language, literature, and traditions of the Chinese nation and has helped to maintain political stability by keeping open to every citizen the door to prestige and power. On the other hand, it has often produced more graduates than could be given positions, has offered little assurance that the successful candidates possessed the qualities necessary for good officials, has sometimes resulted in corruption in the conduct of the examinations, and has in some degree stifled progress.

Some kind of measurement or evaluation seems inevitable in education. It seems inherently an essential part of the teaching process.¹ The situation has been well expressed as follows:²

As far back as we have any record of school routines, teachers have always examined or tested, as well as taught. But our attitudes towards these two functions have been, historically, quite different. We have long understood that teaching is a highly skilled business, a profession calling for special aptitudes and extensive preparation, and that both its techniques and its objectives are worthy of the most careful investigation. But examining or testing we have taken for granted as something that anybody could do any time, quite casually, for any purpose he might happen to think of. It is only yesterday that it occurred to most of us that there might be skilled techniques of testing or that the uses we were accustomed to make of our tests and examinations might be open to question.

Every teacher or administrator of more than twenty years' service will recall with me that Age of Innocence when a "test" regularly consisted of ten questions, sometimes concocted impromptu as we wrote them on the blackboard, each weighted, by our arbitrary personal fiat, with a value of 10 on a scale of 100, and when the perfectly simple purpose of any "test" was to "pass" or "flunk" the testees. We knew no qualms in those days about reliabilities or validities or comparability and the sigma lay as far in the future (for us teachers) as television.

It was only after the World War that this primal innocence was disturbed by the coming—into the consciousness of teachers generally, as distinguished from the psychologists—of what many of us still think of as "the new tests." A bewildering series of strange inventions: intelligence tests first and then objective achievement tests and aptitude tests and interest tests, and personality inventories and ratings. Nearly all of them appallingly elaborate and alleged to have been most laboriously

¹ In this connection the experience of Russia is instructive. In 1917 the Soviet Government wishing to achieve as complete a transformation of education as of government did away with all forms of examinations and school marks. After fifteen years' experience however the Central Committee of the Communist Party declared the plan ineffective and undesirable and recommended the reintroduction of a rigid system of examinations and marks. See Earnest Martin Hopkins, "Prerequisites of Intelligence," *School and Society* 40: 473-480 October 13 1934, "Changes in the Soviet Educational System," *School and Society* 42: 836-837 December 14 1935 and "Testing Russian Students," *School and Society* 50: 25-26 July 1 1939.

² Max McConn, "Examinations Old and New: Their Uses and Abuses," *Educational Record* 11: 375 October, 1915.

prepared, with every item studied and checked and to have been tried out on hundreds or thousands of students, and then re-studied and re-checked by mysterious statistical methods

The foregoing picturesque statement is accurate enough for "teachers generally," as the author intended but is far from true of certain outstanding leaders in the profession. Horace Mann,⁵ for example almost one hundred years ago, had a remarkable conception both of the importance of examinations and of the limitations of the forms then in existence. His penetrating analysis of the weaknesses of the oral examinations then in vogue, and of the superiority of written examinations, could hardly be improved upon by the modern specialist in measurement. Mann showed clearly the points where the oral examinations were lacking, in the technical language of today in validity, reliability, and usability.⁶

Another American educator who understood both the value and the limitations of examinations was Emerson E. White widely known as a writer and school administrator. In 1886 he wrote "It may be stated as a general fact that school instruction and study are never much wider or better than the tests by which they are measured."⁷ In the same volume the author enumerates several "special advantages" of the written test.⁸

It is more impartial than the oral test since it gives all the pupils the same tests and an equal opportunity to meet them, its results are more tangible and reliable it discloses more accurately the comparative progress of the different pupils in formation of value to the teacher, it reveals more clearly defects in teaching and study, and thus assists in their correction, it emphasizes more distinctly the importance of accuracy and fullness in the expression of knowledge it reveals more fully than the ordinary language exercise the ability of the pupil to write correctly when his attention is directed to the thought or subject-matter it is at least an equal test of the thought-power or intelligence of pupils since this result in both methods is dependent upon the nature of the tests, and lastly the certainty of the coming written test affords a healthy stimulus to pupils increasing their attention to instruction, and their efforts to master the subjects taught

These views of Mann and White appear surprisingly modern and show how far the practice of the rank and file is likely to fall behind the theory of the pioneer thinker. It is doubtful if any single sentence in recent educational literature states the superiority of the written over the oral examination more completely or more forcefully than the one just quoted from

⁵ Otis W. Caldwell and Stuart A. Courtis: *Then and Now in Education 1840-1923* pages 37-41. Yonkers: World Book Company, 1923.

⁶ These terms will be explained in Chapter 4.

⁷ Emerson E. White: *The Elements of Pedagogy*, page 148. New York: American Book Company, 1886.

⁸ *Ibid.*, pages 197-198. However in an unusual article dealing with a semi-objective use of oral questioning in the classroom Max M. Koestick and Bille M. Nixon present another side of the argument. "How to Improve Oral Questioning." *Psychology Journal of Education* 30: 201-217 January 1933.

White In fact, many modern specialists in measurement would probably accept the above indictment of oral tests *in toto*. They would, of course, wish to discount somewhat the values so enthusiastically proclaimed for ordinary written examinations, and would point out that many of the limitations of the oral tests so forcefully stated also hold in some degree for the written tests, and in addition that the latter have some special limitations of their own not then recognized. But that is another story to be told later.

B. The History of Intelligence Tests

In Jevons' *The Principles of Science*, published in 1874, occurred this significant statement ⁹

As physical science advances, it becomes more and more accurately quantitative. Questions of simple logical fact after a time resolve themselves into questions of degree, time, distance, or weight. Forces hardly suspected to exist by one generation, are clearly recognised by the next, and precisely measured by the third generation. But one condition of this rapid advance is the invention of suitable instruments of measurement. Accordingly the introduction of a new instrument often forms an epoch in the history of science.

While the foregoing statement was intended as a history of the past development of physical science, it is also a remarkably accurate prophecy of the future development of measurement in psychology, which Jevons appears to have foreseen, as indicated by his reference to the "fact that man in his economical, sanitary, intellectual, aesthetic, or moral relations may become the subject of exact sciences, the highest and most useful of all sciences" ¹⁰. This statement is all the more remarkable when one considers that it was made five years before Wundt established the first psychological laboratory and Galton began publishing his most important studies of individual differences, while both Binet and Cattell were lads in their teens, and before either Thorndike or Terman had been born. But it was over a quarter of a century before any very definite progress was made toward fulfilling the prophecy. Then there followed rapid progress in that direction along several lines. That story will now briefly be told.

Germany and experimental psychology. An important event in the history of psychology was the establishment of the first experimental laboratory in psychology by Wilhelm Wundt at Leipzig in 1879. He was, however, primarily interested in the analysis of consciousness into elements in a manner analogous to that employed in atomic chemistry. His sole interest in measurement appeared to be confined to reaction times, and he was distinctly unsympathetic to the problem of individual differences, but he did influence considerably the course of psychology, especially the work of other German psychologists, such as Kraepelin, Fbbinghaus, and

⁹ W. Stanley Jevons, *The Principles of Science*. Book III, page 313. New York: The Macmillan Company, 1874.

¹⁰ *Ibid.* page 395.

Meumann, who introduced many forms of separate tests, which were borrowed by later investigators in constructing their scales for measuring general intelligence. Of the test forms, the completion test of Ebbinghaus was doubtless the most important.

Another important idea, suggested in 1912 by Stern, was that of representing intelligence as the ratio of mental age to chronological age. This concept, for which Stern suggested the term "mental quotient," was later adopted by Terman as the familiar IQ.

England and statistical methods. The distinctive contribution of the English to the measurement of intelligence has been that of statistical methods as a tool for the analysis of test results. Sir Francis Galton, one of the most brilliant and versatile men of the nineteenth century, was the first to treat seriously the problem of individual differences in psychology, particularly in the realm of sensory discrimination, although Weber, Fechner, Helmholtz, and others had given slight attention to it in what is often termed psychophysics. In 1883 Galton outlined a method for studying free association by quantitative methods. But his most notable contribution was in statistical analysis, where he suggested among other things a graphical method of representing correlations.¹¹ Karl Pearson, a pupil of Galton, and Charles E. Spearman still further advanced the science of statistics. Spearman developed his well-known two-factor theory of intelligence on the basis of statistical analysis. Cyril Burt, who has been a leader in introducing and adopting Binet's work in Great Britain, was in 1913 officially appointed school psychologist, possibly the first person in the world to occupy that position.

France and abnormal psychology. The French have long been leaders in abnormal psychology. Consequently, they approached the problem of measuring intelligence from the standpoint of the classification and treatment of the mentally defective. This brings us to the most important name in the history of intelligence testing, Alfred Binet.

It would be hard to find a man who better illustrates Jevons' description of the methods of the genius than Binet. Jevons says:¹²

It would be a complete error to suppose that the great discoverer is one who seizes at once unerringly upon the truth, or has any special method of divining it. In all probability the errors of the great mind far exceed in number those of the less vigorous one. Fertility of imagination and abundance of guesses at truth are among the first requisites of discovery.

This reads as if it were designed specifically to describe Binet, and yet it appeared twenty years before Binet founded *L'Année Psychologique* and thirty years before his first scale for measuring intelligence. He first studied law, then medicine, and afterward worked in a biological laboratory. Later

¹¹ David G. Ryans, 'Francis Galton's Statistical Contributions,' *School and Society* 48: 312-316, September 3, 1938.

¹² W. Stanley Jevons, *op cit*, Book IV, page 221.

he turned psychologist, first of the arm-chair variety, and finally ended as an experimentalist. Furthermore, in an effort to devise a suitable method of measuring intelligence he tried out various head measurements, physiognomy, graphology, and palmistry, before hitting upon the correct approach. Binet never seemed to be quite sure what he meant by "intelligence" what he was trying to measure, for he changed his definitions repeatedly. It is clear, therefore, that he did not hit "at once unerringly upon the truth" and that he did possess to a marked degree "fertility of imagination and abundance of guesses." Such errors as he made, and they were numerous, were not the unintelligent ones of blind trial and error but rather the intelligent errors of judgment, made by acting upon the course which seemed most promising from a survey of the best available facts at hand.

It is doubtless true, as Boring suggests¹²

At close view the course of science seems discontinuous, all at once a "genius" makes a discovery or formulates a theory and productive research follows on immediately. At the greater range of historical perspective, the course of science seems to be continuous and the "genius" appears as an opportunist who takes advantage of the preparation of the times.

Opinions will differ regarding the appropriateness of the word "opportunist" in Binet's case, but there can be no doubt that he did take "advantage of the preparation of the times." Both for his ideas and actual test materials he drew freely from others, notably his fellow countryman, Bin, and his contemporaries in Germany. Nevertheless, Binet did something the others had not done, he began where they left off and continued with a definite contribution both to the theory and practice of testing. On the theory side he enlarged the prevailing concept of intelligence, introducing such ideas as those of judgment, adaptation, and self-criticism. Terman¹⁴ argues that Binet's outstanding contribution to psychometrics was his abandonment of any attempt to measure "intellectual faculties as such." To practice he contributed a technique of scale construction and a finished scale consisting of test situations selected according to predetermined criteria and standardized. The date 1905 is important, therefore, because it marked the appearance of the first scale for the measurement of intelligence, which, crude as it was, has served as the pattern for subsequent tests and scales the world over. The 1908 revision was a definite improvement, and is especially notable for the introduction of the *mental age* concept. Further experimental work resulted in the scale of 1911, the year of Binet's death.

¹² Edwin G. Boring, *A History of Experimental Psychology*, page 452. New York: The Century Company, 1929. Used by permission of Appleton Century Crofts, Inc.

¹⁴ Lewis M. Terman, "The Revision Procedures," page 6, Chapter I in Quinn McNemar, *The Revision of the Stanford Binet Scale*. Boston: Houghton Mifflin Company, 1942.

America and applied psychology. The scene now shifts to America, where the outstanding name is J. McKen Cattell, who was a pioneer along many lines and a promoter of the first rank. More than anyone else, Cattell was responsible for giving to American psychology its practical bent, for with him the practical took precedence over the philosophical. As early as 1885 he began to publish important articles on reaction times and individual differences. It was Cattell¹⁵ who in 1890 suggested the term "mental tests," which was to become a sort of trade-mark for the whole measurement movement. But Cattell was too close to Wundt's laboratory to escape altogether the views of its master. Cattell, therefore, just as did Galton, confined his tests largely to the simpler mental processes, such as sensory discrimination, where individual differences are least, rather than to the higher mental processes, where they are greatest. In other words, both Galton and Cattell attempted to measure intelligence, but with the wrong tools. Very little attention was given either to reliability or to validity. Consequently, in 1901, when Wissler¹⁶ published his analysis of Cattell's tests used with college students, in which he applied for the first time the Pearson correlation technique to test scores, and in which he found little more than chance relationship either among the tests themselves or between the tests and college work, a considerable damper was thrown over the enthusiasm of American testers which was not lifted till after Binet had published his 1905 scale. Nevertheless Cattell's influence upon measurement, through both his writing and his students, notably Thorndike, has been great.

Goddard was probably the first American psychologist to recognize the practical value of Binet's 1908 and 1911 scales, which he translated and with minor adaptations tried out at Vineland.¹⁷ In 1911 and 1912 Kuhlmann published his revisions of the 1911 Binet scale, extending it downward to the age of three months, instead of three years, which was Binet's lower limit.¹⁸

It remained for Terman of Stanford University to provide the first thoroughgoing revision, carefully adapted to and standardized for use with American children, normal as well as subnormal. Terman's scale, known as the Stanford Revision or Stanford-Binet, appeared in 1916, together

¹⁵ J. McK. Cattell, "Mental Tests and Measurements," *Mind* 15: 373-380, 1890.

¹⁶ Clark Wissler, "The Correlation of Mental and Physical Tests," *Psychological Review*, Monograph Supplement, Vol. VIII, No. 16, 1901.

¹⁷ Henry H. Goddard, "Four Hundred Feeble-minded Children Classified by the Binet Method," *Pedagogical Seminary*, 17: 387-397, September 1910, "A Measuring Scale for Intelligence," *The Training School*, 6: 146-155, January 1910, "Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence," *Pedagogical Seminary*, 18: 232-259, June, 1911.

¹⁸ For an excellent historical discussion, see Florence L. Goodenough, *Mental Testing, Its History, Principles, and Applications*, Chapter IV, "The Early Tests (1887-1915)," and Chapter V, "Later Developments." New York: Rinehart & Company, Inc., 1919.

with a most complete manual, *The Measurement of Intelligence*¹⁹ This revision has been criticized on the ground that it was standardized entirely on school children, which may result in somewhat of a handicap for those of poor academic background, and that it did not produce a sufficient "scatter" in the distribution of IQ's, particularly at the higher ages It has also been criticized on the ground that its norms were based exclusively on the children of one state, California, which may not be truly representative of the United States as a whole Nevertheless, the Stanford Revision was, for more than two decades, the most widely used and most highly regarded individual intelligence test in existence In 1937 a thorough revision of the Stanford-Binet appeared²⁰ This second revision corrected most of the weaknesses of the first revision

Two other distinctly American developments, both aiming to make intelligence tests more practical, remain to be discussed These early tests had two practical disadvantages which militated against their wide use One of these was that the tests were highly *verbal*, that is, their successful administration required that the subject taking the test understand the English language The other was that the tests were *individual* that is, only one person could be examined at a time Reasonably satisfactory solutions came to both of these problems in the year 1917, and this leads to an interesting story

Intelligence tests, children of necessity. One cannot but be impressed with the curious role of necessity in the development of intelligence testing both in Europe and in America Although it may be true, as Thorndike suggests, that necessity is not the true mother of invention, she is, often at least, the stern, relentless stepmother Two instances in Europe and two in America will suffice to make this clear

In 1897 Ebbinghaus was appointed on a commission to investigate the problem of fatigue in the schools of Breslau As there were in existence no appropriate tests, Ebbinghaus set about to devise them, the first "completion tests" resulted from this endeavor Seven years later the Minister of Public Instruction in Paris became concerned about the high percentage of failure in the Paris schools and appointed Binet on a commission to determine those who were so mentally unfit as to necessitate instruction in special classes Binet, too, found available measuring instruments inadequate for the purpose²¹ Out of this difficulty emerged the 1905 scale already referred to, the first successful instrument for measuring intelli-

¹⁹ Lewis M. Terman *The Measurement of Intelligence*, 362 pages Boston Houghton Mifflin Company, 1916

²⁰ Lewis M. Terman and Maud A. Merrill *Measuring Intelligence*, 461 pages Boston Houghton Mifflin Company, 1937

²¹ Binet confessed himself unable to distinguish an idiot who was described by existing standards as having a 'gleam' of intelligence and an attention which was 'fugitive' from an imbecile who was described by these standards as having a 'very incomplete degree' of intelligence and an attention which was 'fleeing' Verily, here indeed was a distinction without a difference

gence according to modern conceptions. But the original Binet scales and their early revisions, both in Europe and America, possessed the two limitations mentioned above, namely, they were highly linguistic and they were individual scales. Soon American ingenuity was to offer solutions to both problems.

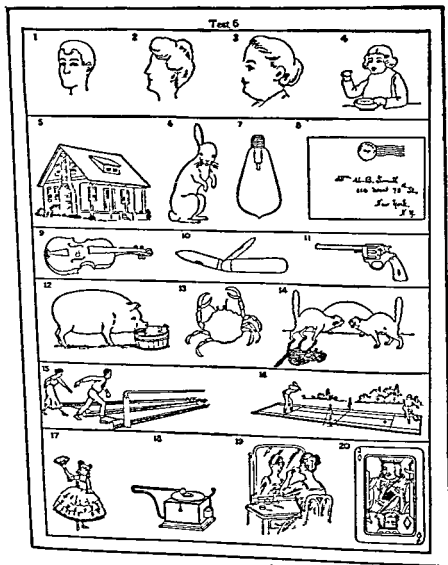
TEST 7	
SAMPLES	sky-blue grass-table green warm big fish-swims man-paper time walks glad day-night white-red black clear pure
In each of the lines below the first two words are related to each other in some way. What you are to do in each line is to see what the relation is between the first two words, and underline the word in heavy type that is related in the same way to the third word. Begin with No. 1 and mark as many sets as you can before time is called.	
1 finger-hand toe—bow foot doll coat	1
2 shirt—hair sleep—book tree bed sea	2
3 shirt—guy trousers—boy hat vest coat	3
4 December—Christmas November—month Thanksgiving December early	4
5 above—day below—above bottom sea hang	5
6 spoon—soup fork—knife plate cup meat	6
7 bird—song man—speech woman boy work	7
8 corn—horse bread—day flour man butter	8
9 sweet—sugar sour—sweet bread man vinegar	9
10 devil—bad angel—Gabriel good face heaven	10
11 Edison—photograph Columbus—America Washington Spain Cuba	11
12 cannon—rifle b g—bullet gun army little	12
13 engineer—engine driver—harness horse passenger man	13
14 wolf—sheep cat—fur kitten dog mouse	14
15 officer—private command—army general obey regiment	15
16 hunter—gun fisherman—fish net hold wet	16
17 cold—heat ice—steam cream frost refrigerator	17
18 aunt—stephew aunt—brother sister niece cousin	18
19 framework—house skeleton—bones skull grace body	19
20 breeze—cyclone shower—bath cloudburst winter spring	20
21 pitcher—milk vase—flowers pitcher table pottery	21
22 blonde—brunette light—house electricity dark girl	22
23 abundant—cheap scarce—costly plentiful common gold	23
24 polite—impolite pleasant—agreeable disagreeable nice face	24
25 mayor—city general—pe cain navy army soldier	25
26 succeed—fail praise—lose friend God blame	26
27 people—house bees—thrive sting have thick	27
28 peace—happy new war—grief fight battle Europe	28
29 a—b c—e b d letter	29
30 darkness—stiffness light—moonlight sound sun window	30
31 complex—simple hard—brittle money easy work	31
32 music—noise harmon cup—beer accord violin discordant	32
33 truth—gentleman lie—rascal live give falsehood	33
34 boy—anger caress—woman kiss child love	34
35 square—cube circle—line round square sphere	35
36 mountain—valley genius—idiot write think brain	36
37 clock—time thermometer—cold weather temperature mercury	37
38 fear—anxiety regret—vacuum memory express road	38
39 hope—cheer despair—grave repair death depression	39
40 dismal—dark cheerful—laugh bright horse slimsy	40

Courtesy National Academy of Sciences

Figure 1. Test 7 from the Army Alpha

Pintner and Paterson, finding the Stanford-Binet unsatisfactory for deaf children, met the first difficulty by assembling a series of fifteen tests of manipulation or performance, such as the form board already used by Seguin, Healy, and others. This combination, which appeared in 1917, was known as the Pintner-Paterson Performance Scale. That same year the United States found herself in World War I faced with the urgent necessity of training a large citizen army with an insufficient supply of commissioned

and noncommissioned officers. In this emergency the American Psychological Association placed its services at the disposal of the War Department. The existing individual intelligence tests were not only entirely unsuited for use with illiterate and foreign-speaking recruits, but they were also



Courtesy National Academy of Sciences

Figure 2 Test 6 from the Army Beta

much too slow. To meet this need, a committee of psychologists, utilizing largely the as yet unpublished work of Otis, prepared the Army Alpha.²² The first of a long succession of group tests destined to receive wide use. The second difficulty of the early tests had now been solved, for a group test can be administered to a hundred or more in the time formerly required for measuring one.

²² Although the Army Alpha had antecedents developed over the preceding 30 years, none of these earlier tests can be said to have passed beyond the experimental stage.

It should be noted, however, that the Pintner-Paterson Performance Scale and the Army Alpha group tests each solved but one difficulty at a time. In fact, as a rule, group tests of the Army Alpha type are even more verbal than the individual tests had been. Figure 1 shows a sample page from the Army Alpha. The early performance scales, on the other hand, were nonverbal, but could be administered to only one person at a time. The Army Beta, designed for illiterate and foreign-speaking soldiers, was the first test to combine the group and performance ideas. It also appeared in 1917. Figure 2 shows the picture completion test of the Army Beta. Since that time several group tests, largely or primarily of the performance type, have been designed specifically for use with young children just entering school. There can be little doubt that World War I gave a decided impetus to the measurement movement in America. World War II had a similar and perhaps even more marked effect.²³

Tests of specific aptitude. All of the tests so far described have been for the measurement of *general* intelligence. There has also been some activity in the development of tests of *specific* intelligence, or capacity in a restricted area, such as music or mechanics, or in a specific school subject, such as algebra or Latin. America has also had the lead in the development of these tests, often called aptitude or prognosis tests. One of the earliest and in many ways one of the best known of these tests is the Seashore Test of Musical Talent, which appeared in 1915. Three years later appeared the Stenquist Test of General Mechanical Ability. In 1918, also, Rogers published a test of mathematical ability, which, although hardly an aptitude test in the modern sense, introduced the idea which other authors have followed up by aptitude tests in the special branches of mathematics, such as algebra and geometry. A somewhat different type of test on the college level is illustrated by the Iowa Placement Examinations which appeared in 1924. A recent and promising type of test, of which there are several examples, is that of reading readiness, to be used to determine a child's fitness for the work of the first grade. There is evidence that the development of the future is likely to be along the line of tests for *specific* aptitude, rather than tests of *general* intelligence, which aim to cover the whole range of human capacity at one shot. The test maker, as well as the bird hunter may aim at too large a target. Dunlap argues this side of the case well: "The more 'general' the intelligence test, the less its value. By increasing the specificity . . . we add to its value. Charles Dudley Warner once shot a bear by 'aiming at it generally,' but it is a poor method." Thurstone's attempt to devise tests of what he terms "primary mental abilities" is a

²³ For a discussion of the Army General Classification Test see Staff Personnel Research Section, "The Army General Classification Test," *Psychological Bulletin* 42 760-768 December, 1945. A civilian edition of the AGCT was published by Science Research Associates in 1947.

²⁴ Knight Dunlap, *Habits Their Making and Unmaking*, page 266 New York Liveright Publishing Corporation 1932.

move in this direction, although these tests have not yet demonstrated marked superiority over other tests in practical schoolroom situations²⁵

C. The History of Achievement Tests

Progress before 1918. The early history of things which have been in existence a long time is usually somewhat obscure. This is true of achievement tests, whose ancient use has already been referred to. Not only have some kinds of tests been in existence for centuries, but attention has been called to the fact that criticisms of them, both destructive and constructive, are by no means new.

But the actual work of improving the existing instruments has always lagged far behind the theory, and actual school practice has been furthest behind of all. In spite of the marked superiority of written examinations over oral, pointed out by Horace Mann in 1845, educators did not forthwith adopt the former or improve the latter.²⁶ However, as early as 1864 an English schoolmaster, the Reverend George Fisher,²⁷ evidently realizing the subjectivity of ordinary examinations, proposed a "Scale-Book," made up of "various standard specimens arranged in order of merit."²⁸ But Ayres observes that "Mr. Fisher's efforts seem to have produced no lasting results," for which he suggested this explanation:²⁹

Progress in the scientific study of education was not possible until people could be brought to realize that human behavior was susceptible of quantitative study, and until they had statistical methods with which to carry on their investigations.

Although Ayres felt that Galton's work had largely met these two needs, he gave Dr. J. M. Rice the honor of being the "real inventor of the comparative test" in America in 1894.³⁰ Rice had studied in Germany and had come under the influence of experimental psychologists both at Jena and Leipzig. Here again the attitude of the educational leaders was anything but cordial, and "for more than ten years but little progress was made beyond the work of the pioneer himself."³¹

²⁵ Duane C. Shaw. A Study of the Relationships between Thurstone Primary Mental Abilities and High School Achievement, *Journal of Educational Psychology* 40: 239-249 April 1949.

²⁶ As Caldwell and Courtis observe. Very few schoolmen proved to have the intelligence of Horace Mann and the era foreseen by him did not begin to materialize until more than fifty years later. *Op cit* page 8.

²⁷ For a good discussion of the early history of achievement tests, see Leonard P. Ayres. History and Present Status of Educational Measurements. *Seventeenth Year Book of the National Society for the Study of Education Part II*, pages 9-15. Bloomington, Illinois: Public School Publishing Company, 1918.

²⁸ Quoted by Edward L. Thorndike and Isaac I. Kandel in Educational Measurement of Fifty Years Ago. *Journal of Educational Psychology* 4: 551-552, November 1913. From E. Chadwick's article in *Museum: A Quarterly Magazine of Education, Literature and Science* 3: 480-481, 1864, where he quotes the Reverend Fisher.

²⁹ Leonard P. Ayres. *Op cit* page 10.

³⁰ Leonard P. Ayres. *Ibid* page 11.

³¹ Leonard P. Ayres. *Ibid* page 12.

Ayres makes a distinction between the "inventor" of educational measurement and the "father" of the movement. The latter distinction he awards to Dr Edward L Thorndike. The honor is richly merited for no other person has touched the measurement movement at so many points or has contributed so much to it. In addition to his very influential publications on statistical methods in education and his pioneer work on intelligence tests for college entrance, either Thorndike or his students were responsible for most of the early standard tests and scales for measuring achievement. The first test was the Stone Arithmetic Test published in 1908 and the first scale was the Thorndike Handwriting Scale announced in 1909 and published the following year. The next few years saw the appearance of scales and tests in various fields. The school survey movement undoubtedly added impetus to the measurement movement, as did the appearance of certain important books and periodicals to be referred to later.

Studies in the unreliability of school marks and examinations. But there was an additional factor which served as a very strong stimulus to standard tests. *Educators discovered for the first time just how bad existing measurements were.* Beginning about 1910, several studies in rapid succession made this point convincingly clear. A distinction should be made between the limitations of school marks and the limitations of school examinations. The need for reform in college marking was forcibly brought to public attention by Max Meyer,³² who reported on the marks collected from forty instructors for a period of five years at the University of Missouri. He found such astonishing variations as 55 per cent of A's in philosophy and only 1 per cent in Chemistry III, while there were 28 per cent of failures in English II and none in Latin I. Johnson³³ found a similar condition in the University of Chicago High School. In a two-year period he found, for example, that the marks for German showed 17.1 per cent A's and 8.4 per cent F's, whereas the marks in English showed 6.5 per cent A's and 15.5 per cent F's. Such variations both at Chicago and Missouri could be most reasonably interpreted on the supposition, not that English is harder than foreign languages but that English instructors are harder. In other words, school marks are highly subjective, the mark received often being more a function of the personality of the instructor than of the performance of the student. Further studies showed similar results elsewhere without exception. This was certainly disturbing if not as Thorndike suggests, actually "scandalous."³⁴

But the evidence presented by a second type of study was even more damaging. Variations among the final marks in different departments

³² Max Meyer 'The Grading of Students' *Science* 28: 243-250 August 21 1908

³³ Franklin W. Johnson 'A Study of High School Grades' *School Review* 19: 13-24

January 1911

³⁴ *Twenty-First Yearbook of the National Society for the Study of Education Part I*

might be accounted for, at least in part, by variations in the background, intelligence, and application of the students in these departments. Thus, at any rate, provided a comfortable loophole. But even this avenue of escape was soon to be closed. Manifestly, such factors could not be responsible for differences when several persons were marking the same student's paper, and least of all when the same person marked the same paper on two different occasions. But studies in abundance have established both conditions.

Perhaps the most striking of the early studies were those of Starch and Elliott. In one of these studies Starch and Elliott³⁵ used facsimiles of the same geometry paper which were marked by 116 high-school teachers of mathematics. The values assigned ranged from 28 to 92. Manifestly, if high-school teachers cannot agree any more closely than that in mathematics, one of the most objective subjects, the situation is indeed bad.

Other studies tended but to confirm the suspicions. One of the most spectacular was made by Falls³⁶ who had 100 English teachers mark a composition by assigning it a percentage value and also indicating the school grade in which they would expect that quality of work to be done.

TABLE 1

THE ESTIMATED GRADE-VALUE AND PERCENTAGE MARKS ASSIGNED TO AN ENGLISH COMPOSITION BY ONE HUNDRED TEACHERS (AFTER FALLS)

Grade-Value	Percentage Mark								Total
	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	
XV								2	2
XIV									0
XIII							1	2	3
XII					1		2	3	6
XI			2			6	5	2	15
X			1	3	8	4	7	1	24
IX	1		1	1	8	4	4	3	22
VIII			2	2	2	3	4	3	16
VII				2	2	2	1		7
VI	1				1	1			4
V	1							1	1
Total	3		6	8	22	20	24	17	100

As will be noted from Table 1, the percentage values varied from 60 to 98, and the estimated grade location, from the fifth grade to the junior year of college. As a matter of fact, the composition was the best one found by a

³⁵ Daniel Starch and Edward C. Elliott, "Reliability of Grading Work in Mathematics," *School Review* 21: 254-259, April 1913.

³⁶ J. D. Falls, "Research in Secondary Education," *Kentucky School Journal*, 6: 42-46, March, 1928.

survey committee at Gary, Indiana, a few years earlier, and was written by a high school senior whose special interest was journalism and who was a correspondent for some of the Chicago newspapers. It is not unreasonable to suppose that many of these English teachers will never have as good a composition submitted by one of their pupils or that few of these teachers could have written a better one themselves.

Evidence is available that examiners in other fields show variations fully as startling as those reported in public education. Ten examination papers written by applicants for licenses to practice dentistry in Kentucky were submitted for regrading to the regular examiners on the official boards of 23 other states. The results are summarized in Table 2. The papers are arranged in rank order from low to high (1 being highest) according to the median judgment of the 24 examiners who are designated by the letters A to X according to degree of strictness in marking. That the variations are enormous is indicated by several facts. With a minimum passing mark of 75, it will be noted that every paper was passed by at least four examiners, and failed by at least four other examiners. The most liberal examiner, A, passed them all while the two strictest W and X failed them all.³⁷ Seven different papers were rated by one or more examiners as the best of the ten while two of these seven papers were rated by other examiners as the poorest of the ten. Surely such a situation can hardly be regarded as anything but chaotic.³⁸

But Starch³⁹ also presented the problem in a different and still more unfavorable light. He found that college instructors assigned different marks when they regraded their own papers without knowledge of their former marks. In a later study Ashbaugh⁴⁰ had 49 Ohio State University seniors and graduate students, the latter with teaching experience, rate a seventh grade arithmetic paper on a percentage basis three times at intervals of four weeks between ratings. Some idea of the lack of consistency in scoring can be gained when it is mentioned that only one student gave the same total score on all three trials and only seven gave the same total score on any two successive trials. The mean differences between pairs of scores on successive trials were as follows: between the first and second trials, 8.1 points; between the second and third trials, 7.3 points. In studies by the writer using the same arithmetic paper, he has found variations of as much as 27 points on successive trials by the same scorer and as much as 10 points variation on values assigned to the first problem on two successive trials approximately ninety days apart.

³⁷ For a fuller account of this investigation see Leon M. Childers, Report of the Research Committee on Examinations, *Proceedings of the Sixtieth Annual Meeting National Association of Dental Examiners* 60-77, 106, August 1942.

³⁸ Daniel Starch, *Reliability and Distribution of Grades*, *Science* 38, 630-631, October 31, 1913.

³⁹ E. J. Ashbaugh, *Reducing the Variability in Teachers' Marks*, *Journal of Educational Research* 9, 185-198, March 1924.

TABLE 2
PERCENTAGE VALUES ASSIGNED TO TEN ESSAY EXAMINATION PAPERS BY TWENTY-FOUR EXAMINERS

Percentage Values Assigned	Rank Order of the Ten Papers (10 is Lowest)									
	Ten	Nine	Eight	Seven	Six	Five	Four	Three	Two	One
100							M			
95 99						(A)C	(A)	(A)F	(A)	FGJ (A)M
90 94					B	GKO	BCEGHIN	BD	BCDEFHM	BCDEHIOQR
85 89				(A)BH	(A)CDEFM	DFHNPQR	DFJKO	CEHJLMNOQS	GIJQR	LN
80 84		(A)CEH		CDEF	GJ	EMT	LPQSU	GR	KOST	KPST
75 79	(A)DKL			GILMNPV	HILNORS		RT			
70 74	ENT	BDCKLMR	FKOQRT	RST	KPQTV	BJL		T	MNU	U(W)
65 69	INOPRS	JNOPST	MNP	JOU	U(X)	S(W)	V	UV	V(W)	V
60 64	CC(W)	IQ		Q		U				
55 59	B			(W)(X)		V(X)	(W)	(W)	(X)	(X)
50 54	F	U(W)					X			
45 49	JQUV	V								
40 44										
35 39	N									
30 34										
25 29										
20 24	(X)									
Med an Range	66 22 75	70 35 83	73 39 96	77 55 87	78 47 90	80 51 91	82 45 100	82 59 92	83 50 90	87 62 98

In a similar study, Hulten⁴⁰ found that 28 experienced Wisconsin high school English teachers differed widely on trials at an interval of two months in the values assigned an English composition which they thought was written by an eighth grade pupil but which was really part of the Hudelson Scale at the time new and unfamiliar. He found that 15 teachers who gave passing marks the first time would have failed the pupil the second time the paper was marked and that 11 teachers who gave failing marks the first time would have passed the pupil the second time. Studies involving English composition are especially significant because every essay examination is a series of compositions and when English teachers who presumably have more than ordinary skill in this field can agree neither with other teachers nor with themselves a second time the situation is very serious.⁴¹

In February, 1918 Thorndike published what has proved to be probably the most influential paper that has ever appeared on educational measurements. It began with the well known dictum "Whatever exists at all exists in some amount," and ended with this note of satisfaction "Of the gains made in the past decade, we may well be proud." As he looked into the future, Thorndike saw it conditioned by a series of ifs.⁴²

If those who object to quantitative thinking in education will set themselves to work to understand it if those who criticise its presuppositions and methods will do actual experimental work to improve its general logic and detailed procedure if those who are now at work in devising and in using means of measurement will continue their work the next decade will bring sure gains in both theory and practice.

We shall now take a look at what really happened in the years following Thorndike's statement of possible achievements.

Progress since 1918 According to Buckingham,⁴³ it was in 1919 that "test-making passed from an amateur to a professional basis." A good summary of the next decade has been made by Monroe⁴⁴ to which reference has already been made. The monograph begins with the assurance that the pioneer state of educational research is passed and that "quantity production" has been achieved. And, as is to be expected, much of the output is

⁴⁰ C. E. Hulten, "The Personal Element in Teachers' Marks," *Journal of Educational Research* 12:49-55, June 1925.

⁴¹ Nor is the situation peculiar to America. Studies reported in Europe reveal conditions fully as bad. In England, for example, examiners were found to change their judgments considerably when they were asked to mark again the same papers they had scored a year before. See *School and Society* 44:364, September 19, 1936.

⁴² Edward L. Thorndike, "The Nature, Purposes, and General Methods of Measurements of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education, Part II*, 1918, pages 16-21. Quoted by permission of the Society.

⁴³ R. B. Buckingham, "Our First Twenty-five Years," *Proceedings of the National Education Association* 1941, page 351.

⁴⁴ Walter S. Monroe and others, *Ten Years of Educational Research, 1918-27*, 309 pages, Bureau of Educational Research Bulletin No. 42, Urbana: University of Illinois, 1928.

not up to the highest quality, when judged by modern standards Monroe, however, detects some evidence of a growing conviction that the emphasis should be upon quality of work rather than upon mere quantity

Moreover, by 1927 there were already developments in new directions which represented a distinct advance The earlier standard tests of achievement were largely of the general or *survey* type, which afforded a general all around measure of the pupil's attainment in the subject, but which did not give the detailed information required for remedial work The next decade saw the development of various achievement tests of a *specific* type For example, there appeared in several fields *diagnostic* tests, whose function was to give specific information regarding the pupil's strong and weak points Also *practice* tests were developed, especially in arithmetic, whose primary function was not so much measurement as drill Another important development of this period was the organization of tests into batteries made up of survey tests in the more important subjects, all published in a single booklet In 1920, two such batteries appeared, one by Pintner and the other by Monroe and Buckingham Two years later appeared the first edition of the well known Stanford Achievement Test, which, with successive revisions, has continued to set a high standard

There was also a rapid development of high school tests in the major academic subjects Even today, however, measurement in high school can hardly be said to have kept pace with that in the elementary school There has also been some activity, but less marked, in the development of achievement tests on the college level

There still remained at the end of the first decade of standardized tests an important need that had not been met Confidence in the ordinary school examination had been seriously undermined by such studies as those to which reference has already been made, and as yet no suitable substitute had been found Also, there were many fields, especially in high school and college where there were hardly any standard tests Even in the subjects most fully provided with such tests, they were by no means adequate to supply the needs of the classroom teacher Furthermore, standard tests represented a considerable item of expense which school boards at that time were often reluctant to assume The so-called *objective*, or *new-type*, test was devised to meet just this need McCall¹ seems to have been the first to suggest this type of test which was merely an adaptation by the classroom teacher of the form of the test items used in the standard test Such tests were usually mimeographed, but they were not standardized Soon they were widely and often uncritically used

Monroe² has given a brief summary of the measurement movement for

¹ William A. McCall, A New Kind of School Examination, *Journal of Educational Research* 13: 16 January 1920

² Walter S. Monroe, Educational Measurement in 1920 and in 1941, *Journal of Educational Research* 35: 334-340 January 1942

the quarter of a century beginning in 1920. It was during these eventful years that educational measurement passed from early adolescence to early adulthood.

Improved examinations, children of necessity. Attention was called earlier in the chapter to the role of necessity in the development of intelligence tests. Much the same influence is evident in the development of improved measurement of achievement. The origin of the objective test referred to above is a case in point. Three other instances will be cited briefly.

It was customary in the early days for the school committees in Massachusetts to give oral examinations in the schools under their control. By 1845 the enrollments had become so large in Boston that the committee could no longer devote the time required for anything more than the most casual examination of each pupil with an oral quiz. To meet this situation the uniform written examination was adopted. The results were so gratifying that Horace Mann wrote the enthusiastic defense of written examinations to which reference has already been made.

In the latter part of the last century considerable pressure was being brought to bear from the outside upon the schools to make place for certain new and practical subjects such as manual training and home economics. But the school men opposed the move on the ground that there was hardly time to teach the subjects already in the curriculum. Then, in 1894 Dr. J. M. Rice had what he called an 'inspiration.' He says: "I

In truth, however, I came to recognize that this (the claims of school men following different courses of study) was all talk -- that no one really knew the facts because there were no standards to serve as guides. Then one day the idea flashed through my mind that the way to settle the question was to try it out. For a beginning I decided to take spelling, and on that very day I made up a list of 50 words with the view of giving them as a test to the pupils of the schools as I went on my tour from town to town. I have no record of the date of the inspiration, but I think it was some time in October, 1894.

This was the origin of the important spelling inquiry which started a movement that not only transformed the teaching of spelling but brought to the fore a new technique for settling educational issues.

The schools of every period have apparently had to meet the criticism that they are not so efficient as those "in the good old days." Usually again there is no defense except argument based upon mere opinions. The criticism was especially severe in the early years of the present century. Just at this time, in 1906, a fortunate event occurred which taught educators a second lesson in the value of comparative examinations. John L. Riley of Springfield, Massachusetts, discovered in an old attic a set of examinations which had been given in the Springfield schools in the year 1846. The thought occurred to him to give these same examinations to the pupils."

"Quoted by Leonard I. Ayres *op cit* page 11. Quoted by permission of the Society

the same city in 1906, just sixty years later.⁴⁸ In spite of the changes in the content of the subjects, he found the results distinctly favorable to the later schools. In ninth grade spelling, for example, the pupils in 1846 had averaged 40.6 per cent, while the average was 51.2 per cent in 1906. In like manner, the geography average had risen from 40.3 per cent in 1846 to 53.4 per cent in 1906. But the greatest superiority was in the case of arithmetic, where the increase was from an average of 29.4 per cent in 1846 to 65.2 per cent in 1906. It was evident, therefore, that the facts were the most effective tools with which to meet criticism, and that comparative examinations were very useful in supplying these pertinent facts.⁴⁹

D The History of Character, Personality, and Interest Measurement

Crude beginnings. It is probably true that human beings began to pass judgment upon each other and to attempt evaluation of each other's character and personality long before the dawn of recorded history. But from the standpoint of measurement, these early efforts were both unsystematic and untrustworthy. Even when somewhat later these methods were reduced to systems the results were still little better than chance. Examples of such prescientific systems which have exerted wide influence upon a credulous public are astrology, graphology, palmistry, and phrenology.

In spite of the antiquity of these attempts at evaluating personality and character, the scientific study of this field is comparatively new. Nor can it be said that the earlier pseudo-scientific approaches have ceased to influence the popular mind. Galton pioneered here as in so many other aspects of measurement. More than 60 years ago he came to the conclusion that the character which shapes our conduct is a definite and durable 'something' and that it is therefore reasonable to attempt to measure it.⁵⁰ He proposed rating scales with statistical analyses of results, and what he termed 'rude experiments' suggested many later investigations. Without doubt Galton's ingenious suggestions marked the beginning of the scientific measurement of character.

Further development. In later years analysis and measurement of personality and character were greatly stimulated by the interest in

⁴⁸ John L. Riley *The Springfield Tests* Springfield, Massachusetts: The Holden Patent Book Cover Co. 1908.

⁴⁹ The discerning student may have noted that the validity of this technique depends greatly upon the obviously erroneous assumption that educational influences *outside* the school were constant from 1846 to 1906. Attempts to prove the effectiveness of modern educational practices by this method usually impress the lay public but from the standpoint of strict logic they are doomed to be inconclusive. In Riley's study, however, the marked superiority found for arithmetic as compared with geography may partially nullify this type of objection.

⁵⁰ Francis Galton *Measurement of Character* *Fortnightly Review* 42: 179-185.

educational and vocational guidance, mental hygiene, and character education.⁵¹ It was soon recognized that success along these lines was conditioned upon the ability to measure other things besides general intelligence and academic achievement. With respect to character education, for example, one of the leaders in that field, Lentz, said "Character education without character measurement would appear to be as logical as target practice in the dark, good shots and poor ones being equally gratifying."⁵²

The first attempt to measure character by a test was probably that of Fernald in 1912, but the author's claims for the test were very modest.⁵³ Voelker, in 1921, devised some actual test situations for measuring character. By far the most ambitious attempt so far made is that of the Character Education Inquiry,⁵⁴ under the direction of Hartshorne and May, which extended over the five-year period, 1924-1929. These workers subjected all the promising tools then in existence to rigid trial and devised many new and ingenious techniques of their own. Their main effort was directed at selecting representative and varied life situations which would afford a valid index of the totality of the character of the individual.

Most of these methods, however, had had interesting historical antecedents. For example, the celebrated physician Galen who lived in the second century, employed methods which were not unlike those in current use. On one occasion Galen, employed by the emperor to find out whether the empress was in love with a certain courtier, attempted to do so by having the suspected lover appear in the presence of the empress while the physician felt her pulse to determine the change in heart beat.⁵⁵ It will be noted that Galen's technique suggests modern physiological methods of blood pressure, psycho-galvanometer, and the like, as well as the "sampling" method of the performance tests.

It is doubtful whether, as a rule, tests of actual performance have sufficiently demonstrated their superior validity and reliability as measures of character and personality to justify the additional expense and inconven-

⁵¹ Some idea can be had of the extensive literature on the subject from examining the annual volumes of the *Review of Educational Research*. A selected bibliography on character including 282 titles selected from a complete list of about 1 000 titles appeared in 1932. A supplementary bibliography for the next three years which appeared in 1935 included over 400 titles. Analogous chapter titles in the 1953 *Educational and Psychological Testing* issue are: Development and Applications of Nonprojective Tests of Personality and Interest and Development and Applications of Projective Tests of Personality. Character measurement per se receives very little attention today. However see Vernon Jones "Character Development in Children—An Objective Approach" Chapter 14 in Leonard Carmichael (Editor) *Manual of Child Psychology* New York John Wiley & Sons 1946.

⁵² Theodore F. Lentz Jr. *An Experimental Method for the Discovery and Development of Tests of Character* page 2 New York Bureau of Publications Teachers College Columbia University 1925.

⁵³ G. G. Fernald "The Defective Delinquent Class Differentiating Tests" *American Journal of Insanity* 68: 524-594 1912.

⁵⁴ A complete report was published by The Macmillan Company in three volumes.

ience involved in their administration, except for purposes of research. In his summary Goodwin Watson says "It is probable that the last five years have brought some swing of psychological interest away from personality-test techniques and toward more emphasis upon the study of personality through ratings anecdotal records, observation of behavior and case studies"⁵⁵ In the final chapter of Symonds' comprehensive and analytical book, *Diagnosing Personality and Conduct*, the author makes this statement "Probably the greatest usefulness will be found in ratings the questionnaire, and the interview for obtaining evidence as to adjustments toward the environment, personal evaluation, attitudes toward reality, sexual relationships, morals, and feelings"⁵⁶ The same author notes clear evidence of a lag both in clinical research and in the translation of this research into educational practice. He says⁵⁷

Work which had been done in the decade from 1910 to 1920 on mental and achievement tests was being assimilated in educational practice in the 1920's. Basic work on the measurement and evaluation of personality that had been carried out in the 1920's was being assimilated in educational practice in the 1930's. Basic work on child guidance procedures, discussions of the meaning of mental hygiene in education and the problems of pupil adjustment which were being investigated and elaborated in the 1930's is only now being assimilated in the practice of education in the schools of the nation.

The historical development of rating scales, questionnaires, and interviews will now receive brief consideration. Strictly speaking, rating scales and questionnaires are only devices for recording the judgments of observers, rather than true measuring instruments.⁵⁸

Rating scales. The first rating scale in a modern sense was probably that of Galton for mental imagery, which was published in 1883. About the time of the appearance of the first Binet test, Karl Pearson proposed a scale for judging intelligence. One of the most famous scales specifically for measuring traits of personality is the Scott Man-to-Man Scale, introduced and extensively used during World War I. It remained for Hartshorne and May to restore somewhat the lost prestige of the rating procedure partly by changing the name to "reputation measures," but mainly by improving the technique.⁵⁹

⁵⁵ *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II*, page 76. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1938.

⁵⁶ Percival M. Symonds, *Diagnosing Personality and Conduct*, page 567. New York: D. Appleton Company, Inc., 1931. Used by permission of D. Appleton Century Company.

⁵⁷ Percival M. Symonds, "The Lag in Clinical Research," *Journal of Educational Research* 38: 371-374, January, 1945. Page 373.

⁵⁸ B. Othanel Smith, *Logical Aspects of Educational Measurement*, Chapter I. New York: Columbia University Press, 1938.

⁵⁹ A historical note by the authors offers another illustration of the kinship between

The questionnaire. The invention of the much-maligned questionnaire is often ascribed to that versatile Englishman, Sir Francis Galton. But that the instrument was in existence at an earlier period in England, in fact if not in name, is evident from the following critical statement:⁶⁰

It is impossible to expect accuracy in returns obtained by circulars various constructions being put upon the same question by different individuals who consequently classify their replies upon various principles

But Galton undoubtedly improved and used extensively the questionnaire, which was imported to America about 1880 by G. Stanley Hall. The instrument exists in many forms today, but its principal use in education is for measuring adjustment, attitude, and interest.

The Woodworth Personal Data Sheet began in 1917 as a method of measuring the ability of soldiers to adjust themselves to the trying conditions of army life. In 1923 Mathews adapted it for school use. The same year Cady published another revision which has been widely used with children in their teens. Two years later Laird made an adaptation for college use and provided a graphic rating scale for scoring. In 1919 Pressey published his widely used X-O Test, a sort of questionnaire covering a miscellaneous assortment of items having to do with emotionality. The questionnaire has also been used to measure other types of adjustment, such as introversion-extroversion by Marston and others, and ascendance-submission by Allport.

The development of wholesome attitudes has been recognized in recent years as an important objective of education. Since about 1920 much attention has been devoted to the measurement of attitudes of various kinds. Hart's test of social attitudes and interests which appeared in 1923 and Watson's measurement of fairmindedness which appeared in 1925 are good examples. Beginning in 1928 Thurstone has been responsible for important improvements in the units of measurement employed in attitude questionnaires on many subjects. By having his questions scaled by a group of judges, Thurstone has found it possible to secure satisfactory results with fewer items. Figure 3 shows an adaptation of this weighting technique applied to a scale for measuring a pupil's attitude toward high school.⁶¹

necessity and invention. For a while it seemed that rating scales as scientific instruments would be completely discarded. It was necessity that saved the day. While everyone talked about the superiority of objective tests, yet it was soon found that many qualities of character yield only stubbornly and expensively to objective testing. If character and personality studies were to continue, ratings had to be revived. In spite of all their difficulties, snares, delusions, and pitfalls, they are now staging a considerable 'comeback.' " *The Journal of Social Psychology* 1:66 February 1930.

⁶⁰ Quoted from *Journal of the Statistical Society of London*, October, 1839, by Walter S. Monroe and Max D. Engelhart, *The Scientific Study of Educational Problems*, page 40. New York: The Macmillan Company, 1936.

⁶¹ For a discussion of this scale, see H. H. Remmers, G. C. Brandenburg, and F. H. Gillespie, "Measuring Attitude Toward the High School," *Journal of Experimental Education* 2:60-64 September, 1933.

In this case the scale values range from 6 for Item 1 to 10.1 for Item 22. The pupil's score is the median scale value of the items checked.

The close relationship of interest to guidance, whether educational, vocational or personal, has stimulated considerable activity directed toward its measurement. In 1907 G. Stanley Hall published a questionnaire study of the recreational interests of children, which exerted a wide influence

HIGH SCHOOL ATTITUDE SCALE*

Form B

Below is a list of twenty-five statements about school. Place a check mark before each statement with which you agree and leave unmarked those with which you disagree. This test will in no way affect your standing in school.

- _____ 1 School is like a prison
- _____ 2 I have a lot of fun in school
- _____ 3 School is all right
- _____ 4 My teachers always treat me fairly
- _____ 5 I like to go to school to be with other people
- _____ 6 Many of our great men have no high school education
- _____ 7 I hate most school work
- _____ 8 Some things about high school are all right
- _____ 9 High school training develops personality
- _____ 10 High school is a good thing for some people and a bad thing for others
- _____ 11 I would just as soon stay at home as go to school
- _____ 12 All the better class of people have high school education
- _____ 13 The high schools lift the plane of sportsmanship in a community
- _____ 14 Too much money is being spent on high schools for the benefit received
- _____ 15 The high school teaches mostly old useless information
- _____ 16 They won't teach things one really wants to know in high school
- _____ 17 If one has plenty of money it may be all right to go to high school
- _____ 18 I haven't any definite like or dislike for high school
- _____ 19 Any old fogy knows more than a high school graduate
- _____ 20 The kindest and best people I know don't have a high school education
- _____ 21 High school cramps and dwarfs one's personality
- _____ 22 America could not stand as a nation if it were not for our high schools
- _____ 23 Our high schools teach immorality and indecency
- _____ 24 High school training develops high ideals in pupils
- _____ 25 High schools develop loyalty

* Prepared by F. H. Gillette and published by Purdue University

Figure 3 A Scale for Measuring Pupils' Attitudes Toward High School

This pioneer study has been followed by many others, of which the most extensive is probably that of Lehman and Witty, published in 1927. Moore appears to have been the first to use this technique for the measurement of vocational interests in 1921. Two other studies, employing somewhat different techniques, appeared in 1926. One of these, by Miner, offered *paired comparisons*, and the other, by Cowdry, employed a complicated scheme for weighting the scores. The best known of all are the Strong Vocational Interest Blank, which appeared in 1927, and the Kuder Preference Record, which followed in 1939.

Paper-and-pencil personality inventories have been severely criticized in recent years, especially by a prominent clinical psychologist, Albert Ellis.⁶² He speaks out against the more elaborately scored and tediously interpreted instruments of this kind.⁶³

since personality inventories depend in the last analysis on printed questions, and since virtually no one claims for them the advantage of getting at unconscious and semi-conscious material which effective clinical use of interview and projective methods are generally conceded to some extent to uncover it is difficult to see why clinicians should spend considerable time first mastering and then using these inventories.

The interview. One of the oldest forms of obtaining knowledge is the personal interview. It has always been an important tool in the hands of certain professional men, such as lawyers, doctors, and newspaper reporters, but it is used by everybody to some extent. Its chief value in education is probably in diagnosis and guidance.⁶⁴ The interview is used to supplement the ordinary objective evidence about a pupil, such as is afforded by his school record and by a firsthand knowledge of such things as his feelings and point of view. The evidence indicates that the personal qualities of the interviewer are fully as important as the technique employed.⁶⁵

Recent developments. Three promising developments relating to measurement require brief mention. The first of these is the attempt to subject personality to elaborate statistical analysis. Leaders in this movement have

⁶² In 'The Validity of Personality Questionnaires' *Psychological Bulletin* 43 385-440 September, 1946. Albert Ellis and Herbert S. Conrad are somewhat more favorable to them in 'The Validity of Personality Inventories in Military Practice,' *Psychological Bulletin* 45 385-426 September 1948.

⁶³ Albert Ellis 'Recent Research with Personality Questionnaires' *Journal of Consulting Psychology* 17 45-49, February, 1953. Page 48. Quoted with permission of the *Journal of Consulting Psychology* and the American Psychological Association. For a rebuttal see Allen Calvin and James McConnell, 'Ellis on Personality Inventories,' *Journal of Consulting Psychology* 17 462-464 December, 1953.

⁶⁴ Thorndike has described the Stanford Binet as 'an approved systematized and standardized interview.' *The Measurement of Intelligence*, page 1. New York: Bureau of Publications, Teachers College, Columbia University, 1927.

⁶⁵ Valuable suggestions on the interview are found in Marie Jahoda, Morton Deutsch and Stuart W. Cook, *Research Methods in Social Relations: Parts I and II*, Chapters VI, VII and VIII. New York: The Dryden Press, 1951.

been Spearman, Kelley, Thurstone, Flanagan, R. B. Cattell, Eysenck, and Stephenson. The second development is that of controlled observation or time-sampling, which is employed mainly in the child study laboratories. A third development has been in the direction of measuring public opinion.⁶⁶

E Some Important Publications

From the beginning professional journals, books, and other publications have exerted a profound influence upon experimental psychology and the testing movement. Only the most important can be mentioned here.

Professional journals. The value of professional journals is that they keep the professional worker in one area continuously informed of what is going on in his own area and elsewhere. The first psychological journal was *Mind*, founded in England by Bain in 1876. For eleven years it remained the only psychological journal in the English language, and so was the vehicle for most of the important psychological articles both in England and in America. One of the most important articles on measurement during the early years was probably that by Cattell entitled "Mental Tests and Measurements," which appeared in 1890 and contained some significant comments by Galton. The first psychological journal in America was the *American Journal of Psychology*, started by G. Stanley Hall in 1887. It has published many significant articles on measurement and statistics, but doubtless none more important than Spearman's "General Intelligence Objectively Determined and Measured," which appeared in 1904. This was the original formulation of the now well known two-factor theory of intelligence and the beginning of "correlational psychology," which was influential in directing the attention of psychologists from faculties to factors.

Hall also founded *Pedagogical Seminary* (now the *Journal of Genetic Psychology*), which may be regarded as the first journal of educational psychology in 1891. Three years later Binet started *L'Annee Psychologique* which was to be the principal agency for bringing his extensive work on the measurement of intelligence to the attention of the world. The *Teachers College Record*, started in 1900, has published many of Thorndike's important studies, and those of his students and co-workers. The first volume of the *Journal of Educational Psychology*, founded in 1910, con-

⁶⁶ Hadley Cantril (Editor) and Mildred Strunk (Compiler) *Public Opinion 1935-46* Princeton: N. J. Princeton University Press 1951 1191 pages.

George H. Gallup *A Guide to Public Opinion Polls* Princeton: N. J. Princeton University Press 1944 104 pages.

Mildred B. Parten *Surveys, Polls and Samples* New York: Harper 1950 624 pages.

For a critical summary and extensive bibliography see Quinn McNemar "Opinion Attitude Methodology" *Psychological Bulletin* 43 289-374 July 1946. This article aroused considerable controversy in the *Psychological Bulletin*. Leo P. Crespi "Opinion Attitude Methodology" and the Polls—a Rejoinder 43 562-569 November 1946. Herbert S. Conrad "Some Principles of Attitude-Measurement: a Reply to Opinion Attitude Methodology" 43 570-589 November 1946 and Quinn McNemar "Response to Crespi's Rejoinder and Conrad's Reply to Appraisal of Opinion Attitude Methodology" 41 171-176, March, 1947.

tained an article by Huey on "The Binet Scale for Measuring Intelligence and Retardation." Two other journals, *School and Society*, and *Educational Administration and Supervision*, both founded in 1915, have included many reports on the use of tests both for research purposes and for the actual work of instruction and school administration. But probably no journal has been more important than the *Journal of Educational Research*, started in 1920. From the very first issue, which contained McCall's article on "A New Kind of School Examination," to the present time, it has exerted a wide influence upon the measurement movement.

Psychometrika, which began publication in 1936 as "a journal devoted to the development of psychology as a quantitative rational science," carries many articles applying mathematical procedures to measurement problems. On a less mathematical level is *Educational and Psychological Measurement*, begun in 1941. It is particularly valuable for the reasonably well trained guidance worker. Even less technical is the *Personnel and Guidance Journal*, formerly called *Occupations*. Two other publications whose articles frequently concern tests are the *Journal of Applied Psychology* and the *Journal of Consulting Psychology*.

Measurement is such an important topic that practically all professional journals as well as many "popular" magazines, deal with various aspects rather often.

Some important books. Some of the important books are briefly mentioned in chronological order, beginning with the pioneer period, which included roughly the first two decades of the century. In 1904 appeared E. L. Thorndike's *An Introduction to the Theory of Mental and Social Measurements*,⁶⁷ which made available for the first time to American students the statistical techniques necessary for educational research and measurement. Ten years later Truman Kelley's *Educational Guidance*⁶⁸ introduced educational workers to the alluring possibilities of partial and multiple correlations. Two important books appeared in 1916. One of these, Daniel Starch's *Educational Measurement*,⁶⁹ was the first book on achievement tests and the other, L. M. Terman's *The Measurement of Intelligence*,⁷⁰ was the first adequate treatment of intelligence tests in the English language. In 1918 appeared the *Seventeenth Yearbook of the National Society for the Study of Education*,⁷¹ Part II of which treats in some detail the history of the pioneer period in testing, and which gives descriptions of existing tests with suggestions as to their use. But it is probably most famous for containing Thorndike's statement, "Whatever exists at all exists in some amount," which has been accepted as a sort of creed by many workers

⁶⁷ Published by Bureau of Publications, Teachers College, Columbia University.

⁶⁸ Published by Bureau of Publications, Teachers College, Columbia University.

⁶⁹ Published by The Macmillan Company, New York.

⁷⁰ Published by Houghton Mifflin Company, Boston.

⁷¹ Published by Public School Publishing Company, Bloomington, Illinois.

in the field The following year, in 1919, appeared Carl Seashore's *The Psychology of Musical Talent*,⁷² a pioneer study in the measurement of aptitude in a restricted field

Since the beginning of the third decade of the century, the "quantity production" stage referred to by Monroe has been achieved not only in the publication of tests but in books as well Only a few of these can be mentioned here as representative of types In 1922 appeared W A McCall's *How to Measure in Education*,⁷³ a comprehensive and critical book on achievement tests The next year saw the publication of Ben D Wood's *Measurement in Higher Education*,⁷⁴ the first treatise on measurement at the college level In 1924 appeared G M Ruch's *The Improvement of the Written Examination*,⁷⁵ which was the first book wholly devoted to the new-type test The year 1927 was especially productive, for at least five important books on measurement bore that date of publication There were two notable books on intelligence, E L Thorndike's *The Measurement of Intelligence*,⁷⁶ and C E Spearman's *The Abilities of Man*,⁷⁷ each representing a distinct point of view In 1927 also appeared the first two books specifically devoted to measurement in the high school, P M Symonds' *Measurement in Secondary Education*,⁷⁸ and G M Ruch and G D Stoddard's *Tests and Measurements in High School Instruction*⁷⁹ The same year Truman Kelley's critical discussion entitled *Interpretation of Educational Measurements*,⁸⁰ was published The next year, in 1928, appeared another critical volume, Clark Hull's *Aptitude Testing*,⁸¹ destined to become one of the classics in the field of measurement

During the 1930's there appeared numerous books and monographs on the various phases of measurement and their application to the different educational levels Some evidence that measurement was coming of age is afforded by the fact that extensive bibliographies of tests and scales appeared during this decade The first edition of Gertrude Hildreth's *Bibliography of Mental Tests and Rating Scales*⁸² was published in 1933 Three years later came Oscar Buros' *Educational, Psychological, and Personality Tests of 1933, 1934, and 1935*,⁸³ the forerunner of *The Mental Measurements Yearbook's*, the first volume of which appeared in 1938⁸⁴ The publication

⁷² Published by Silver, Burdett and Company, New York

⁷³ Published by The Macmillan Company New York

⁷⁴ Published by World Book Company Yonkers

⁷⁵ Published by Scott, Foresman & Company, Chicago

⁷⁶ Published by Bureau of Publications Teachers College, Columbia University

⁷⁷ Published by The Macmillan Company New York

⁷⁸ Published by The Macmillan Company, New York

⁷⁹ Published by World Book Company, Yonkers, N Y

⁸⁰ Published by World Book Company, Yonkers

⁸¹ Published by World Book Company Yonkers

⁸² Published by The Psychological Corporation New York

⁸³ Published by Rutgers University Press New Brunswick, New Jersey

⁸⁴ Published by Rutgers University Press

of this critical volume marked an important milestone in the history of educational measurement. *The Nineteen Forty Mental Measurements Yearbook*,⁸⁵ *The Third Mental Measurements Yearbook* (1949),⁸⁶ and *The Fourth Mental Measurements Yearbook* (1953)⁸⁷ are Buros' invaluable sequels to these earlier works.

David Wechsler's *The Measurement of Adult Intelligence*,⁸⁸ a comprehensive manual accompanying the first individual intelligence test designed especially for testing adolescents and adults, was first published in 1939. The Wechsler-Bellevue Intelligence Scales have been well accepted, particularly by clinicians. The *Wechsler Intelligence Scale for Children*⁸⁹ emerged in 1949 as a serious competitor of the Revised Stanford-Binet Intelligence Scale in the age range 5-15 years.

Though World War II temporarily interrupted the publishing efforts of most measurement specialists, during the post-war years a veritable deluge of excellent books hit the market in rapid-fire succession.⁹⁰ Among these were Frederick B. Davis' *Utilizing Human Talent*⁹¹ and Dorothy C. Adkins and others' *Construction and Analysis of Achievement Tests*⁹² (1947), Lee J. Cronbach's *Essentials of Psychological Testing*,⁹³ Florence L. Goodenough's *Mental Testing*,⁹⁴ William Stephenson's *Testing School Children*,⁹⁵ Donald E. Super's *Appraising Vocational Fitness*,⁹⁶ and Robert L. Thorndike's *Personnel Selection, Test and Measurement Techniques*⁹⁷ (1949), Frank S. Freeman's *Theory and Practice of Psychological Testing*⁹⁸ and Harold Gulliksen's *Theory of Mental Tests*⁹⁹ (1950), and E. F. Lindquist and others' *Educational Measurement*¹⁰⁰ (1951).

The Goheen-Kavrucek *Selected References on Test Construction, Mental Test Theory, and Statistics, 1929-1949*¹⁰¹ greatly facilitate research, putting at one's finger tips the titles and sources of much of the important material which appeared during those two decades.

⁸⁵ Published by the Gryphon Press, Highland Park, N. J.

⁸⁶ Published by Rutgers University Press.

⁸⁷ Published by Gryphon Press.

⁸⁸ Published by the Williams & Wilkins Company, Baltimore, Maryland.

⁸⁹ Published by The Psychological Corporation, New York.

⁹⁰ Julian C. Stanley, "Five Recent Educational and Psychological Measurement Text-books," *Harvard Educational Review*, 22: 57-61, Winter, 1952.

⁹¹ Published by the American Council on Education, Washington, D. C.

⁹² Published by the U. S. Government Printing Office, Washington 25, D. C.

⁹³ Published by Harper & Brothers, New York.

⁹⁴ Published by Rinehart & Company, New York.

⁹⁵ Published by Longmans, Green and Company, New York.

⁹⁶ Published by Harper & Brothers, New York.

⁹⁷ Published by John Wiley & Sons, New York.

⁹⁸ Published by Henry Holt and Company, New York.

⁹⁹ Published by John Wiley & Sons, New York.

¹⁰⁰ Published by the American Council on Education, Washington, D. C.

¹⁰¹ Published by the U. S. Government Printing Office, Washington 25, D. C., 1950.

F. Some Relatively Recent Tendencies

Test construction. The "quantity production" stage of test construction in America now seems definitely past. The emphasis has turned to quality, although it is still too much to say that a recent copyright date on a test is ample assurance of high merit. Kelley's observation made in 1927 that "the ruts of the test movement are already so deep that there are many who do not see beyond them"¹⁰¹ is still, unfortunately, true. However, test makers as a group no longer unblushingly make the enthusiastic claims for their products that were common a few years ago. Instead there has grown up a more critical and becomingly modest attitude, which is probably the most characteristic feature of the present trend. One alert observer¹⁰² as early as 1920 noted "evidences of the beginnings of a critical attitude toward educational tests."

Another tendency, largely an outgrowth of this critical attitude, is to extend the field of measurement into new areas and to develop new, and usually more *specific*, types of tests. For example, instead of a primary interest in developing tests of general intelligence, the emphasis is upon developing tests of specific intelligence along particular lines. Reading readiness and other aptitude tests are representative of the trend. Even so-called general intelligence tests that have appeared recently often yield several scores which have possible diagnostic value. Increased attention to the reliability, and more particularly to the validity, of tests and individual test items is also a notable trend. The result has been the appearance of new types of test forms and test situations. Along with this has come the realization that standard tests do not fully meet all the needs of measurement, and that in consequence greater emphasis must be placed upon the development of improved techniques for constructing informal teacher made tests and other techniques of evaluation.

Monroe makes the following excellent summary of the situation.¹⁰³

The most significant trends appear to be (1) the attack upon and the consequent discrediting of essay examinations (2) the development of objective tests, and the emphasis upon reliability as the criterion by which measuring instruments were evaluated and (3) the development of diagnostic and prognostic uses. What of the future? Any attempt to project lines of development into the future is attended with uncertainty. But if I interpret correctly current educational writings in this field three trends are indicated: (1) a growing emphasis upon validity and a consequent decreasing emphasis upon reliability as the criterion for evaluating

¹⁰¹ Truman Lee Kelley, *Interpretation of Educational Measurements*, page 16. Yonkers World Book Company, 1927.

¹⁰² Walter S. Monroe, "Educational Measurement in 1920 and 1915," *Journal of Educational Research*, 38, 334-340, January, 1915.

¹⁰³ Walter S. Monroe, "Some Trends in Educational Measurement," *Twenty-Fourth Annual Conference on Educational Measurements*, page 33, Bulletin of the School of Education, Indiana University, Vol. XIII, No. 4, Bloomington, Indiana, Bureau of Educational Research, 1917.

measuring instruments, (2) a decline of the faith in indirect measurement and an increasing emphasis upon direct measurement as a means of attaining satisfactory validity, and (3) a growing respect for essay examinations as instruments for measuring certain outcomes of instruction

Use of tests. A British psychologist has suggested that a new scientific technique seems to go through three stages, as follows ¹⁰⁵

The first is the early state of development when no one except its inventors, is interested in it, and those working on other lines regard it with indifference or suspicion or else think it silly. In the second stage it begins to gain support, and in the third stage everyone wants to use it whether they understand it or not. There is then danger of a fourth stage of disillusionment and this is the time for critical examination.

In the case of standard tests in America the stage of indifference and suspicion, with which Rice's spelling inquiry was met, had largely passed when the first standardized tests appeared during the first decade of the present century. Since that time there have been three rather clearly defined stages, which may be designated as those of curiosity, confidence, and critical caution.

The first stage was that of *curiosity*. In this stage teachers and school officials tried out tests merely because they were something new and because their use gave evidence, if indeed superficial in character, of up-to-date-ness. This attitude tended to die a natural death as the novelty wore off.

The second stage was that of *confidence*, or in some instances that of overconfidence. Standard tests were "swallowed, hook, line, and sinker." Test results were uncritically accepted at their face value. IQ's were naively taken as accurate measures of innate capacity wholly apart from environmental opportunities, and so were as fixed as the laws of the Medes and the Persians ¹⁰⁶. In like manner, achievement test scores were accepted as fully adequate measures of the important outcomes of instruction. If only such tests were objective, they were assumed to be sufficiently accurate for valid comparisons, not only of one school or class with another but also of one pupil with another, or even of one aspect of a pupil's achievement with another aspect of his achievement ¹⁰⁷. There is some evidence that this atti-

¹⁰⁵ J. O. Irwin 'Correlation Methods in Psychology' *British Journal of Psychology* 25: 86-91 July 1934

¹⁰⁶ What happened to intelligence testing following World War I has been described as follows: "As many of the subjects tested were children of school age because Binet's scores gave a good correlation with ability for school work and perhaps because of the relative simplicity and economy of the methods mental testing was oversold and careful psychological work in the field of individual differences still suffers from this effect." Francis N. Maxfield 'Trends in Testing Intelligence' *Educational Research Bulletin* 15: 137 May 13 1936

¹⁰⁷ What happened to achievement testing has been described as follows: "In the widespread use of objective tests at the high school and college levels there is apparent a child like faith in the efficacy of objective tests as instruments for measuring school achievement. A little knowledge has become a dangerous thing." Walter S.

tude is on the decline, although unfortunately it is still found too often in certain quarters

The third stage may be termed that of *critical caution*. While by no means universal this more wholesome attitude, on the whole, characterizes the later phases of the testing movement. Hildreth points out some beneficial results of this change. "A more critical attitude toward intelligence measurement, as the outcome of continued experimentation, has resulted in more authoritative research findings, more sensible and intelligent interpretation of data."¹⁰⁸ This attitude has shown itself with respect to achievement tests and personality measurements as well. The result has been not so much the curtailment of the use of tests as their more critical use and the more cautious interpretation of test scores. On the whole the current emphasis is on the use of standard tests not so much for comparative purposes, as to provide a basis for guidance and remedial instruction. It is increasingly recognized that tests are means and not ends, and that even the best test is but a tool the value of which depends upon the skill and the intelligence with which it is used.

This enlarged and more critical attitude on the part of enlightened school officials has been well stated by Maxfield:¹⁰⁹

In problems of school administration the massed data from intelligence tests will be interpreted by statistical methods. In dealing with the problems of individual pupils the case-study method of the clinical psychologist will prevail. Inventories of personality scales of social adjustment and the like will supplement tests of intelligence. Diagnostic tests will be supplemented by diagnostic teaching. But no synthesis or interpretation will be attempted without knowledge of the pupil's physical condition, his home background, his previous school history, his vocational interests, his social and emotional reactions, and the like. The weight given in this synthesis to scores on intelligence tests will vary with the problem presented. The case study method can be adapted to any philosophy of education and to any educational aims and objectives.

Landquist concludes his excellent forty page chapter, "Preliminary Considerations in Objective Test Construction" with a strong warning:¹¹⁰

If measurement is to continue to play an increasingly important role in education measurement workers must be much more than technicians. Unless their efforts are directed by a sound educational philosophy unless they accept and welcome a greater share of responsibility for the selection and clarification of educational objectives unless they show much more concern with what they measure as well as with how they measure it, much of their work will prove futile or ineffective.

Monroe Hazard, in the *Measurement of Achievement*, *School and Society* 41:48-49 January 12, 1932.

¹⁰⁸ Gertrude Hildreth, *Applications of Intelligence Testing*, *Review of Educational Research* 7:200 June 1937.

¹⁰⁹ Francis N. Maxfield, *op cit* page 110.

¹¹⁰ In *Educational Measurement* page 128 Washington: D. C. American Council on Education 1931.

SELECTED REFERENCES FOR FURTHER READING

- Boynnton, Paul L, *Intelligence Its Manifestations and Measurement* New York D Appleton-Century Company, Inc , 1933 Chapters V and VI
- Cook, Walter W , "Achievement Tests in Walter S Monroe (Editor), *Encyclopedia of Educational Research*, pages 1461-1478 New York The Macmillan Company, 1950
- Cook, Walter W , "What Educational Measurement in the Education of Teachers?" *Journal of Educational Psychology*, 41 339-347, October 1950
- Eric F Gardner, "Development and Applications of Tests of Educational Achievement in Schools and Colleges," *Review of Educational Research*, 23 85-101, February, 1953
- Freeman, Frank N , *Mental Tests Their History, Principles and Applications* (Revised Edition) Boston Houghton Mifflin Company, 1939 Chapters I-VIII
- Moody, Caesar B , "Historical Outline of Concepts of Mental Ability," *Peabody Journal of Education*, 30 194 204, January, 1953
- Odell, C W , *Educational Measurement in High School* New York D Appleton-Century Company, 1930 Chapter II
- Peterson, Joseph, *Early Conceptions and Tests of Intelligence* Yonkers, N Y World Book Company, 1925 320 pages
- Pintner, Rudolph, *Intelligence Testing Methods and Results* (New Edition) New York Henry Holt & Company, 1931 Part I
- Rice, Joseph M , *Scientific Management in Education* New York Hinds, Noble & Eldredge, 1913 282 pages
- Ruch, G M , *The Objective or New-Type Examination* Chicago Scott, Foresman & Company, 1929 Chapters I and III
- Rugg, Harold O , *Foundations for American Education* Yonkers, N Y World Book Company, 1947 Chapter XXIII, "Fifty Years of Scientific Method in Education What Have We Learned? "
- Thurstone, Louis L , and Chave, Ernest J , *The Measurement of Attitude* Chicago University of Chicago Press, 1929 96 pages
- U S Office of Strategic Services, *Assessment of Men* New York Rinehart & Company, 1948 541 pages
- Young, Kimball, ' The History of Mental Testing,' *Pedagogical Seminary*, 31 1-48, March, 1924

3

The Statistical Analysis of Test Results

A General Considerations

The importance of statistics Measurement and evaluation are coming increasingly to depend upon statistical procedures Almost all test manuals discuss central tendency, variability, percentiles standard scores, reliability and validity, usually presuming that the reader understands certain commonly used statistics fairly well Educational literature of all types contains such concepts and statistics they are also mentioned at professional meetings Workers in all fields of education can anticipate heightened emphasis upon statistical thinking and techniques

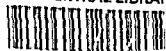
The view taken by Helen Walker is even broader ¹

The conclusion seems inescapable that some aspects of statistical thinking which were once assumed to belong in rather specialized technical courses must now be considered part of general cultural education

Statistics in a capsule Nearly every elementary measurement text, most general psychology books and many other introductory volumes con

¹ Helen M. Walker, "Statistical Understandings Every Teacher Needs," Pages 207-211 in *Improving Educational Research*, Official Report of the American Educational Research Association, Washington, D. C., National Education Association, 1948. Reprinted in *Teachers College Record* 49: 457-467, April 1948 and in *High School Journal* 35: 70-3, January 1950.

Other similar articles are: W. Edwards Deming and Douglas E. Scates, "The Need for Statistical Education in High School and College," *Educational Record* 29: 72-80, January 1945; Douglas E. Scates, "There is a Place for Statistics in General Education," *Youngs School* 41: 47-43, February 1948; Douglas E. Scates and W. Edwards Deming, "Education in Statistics for Participation in Current Affairs," *School Review* 56: 262-269, May 1948; and Millicent Haines, "Thinking Straight about Facts and Figures," *National Leadership* 6: 100-103, November 1946.



tain a chapter on statistics which is hopefully designed to achieve the impossible—that is, to teach in a week or so material usually covered in a quarter, semester, or year-long statistics course. Though some of these chapters are very good indeed, failure to attain their unrealistic goal is frustrating to students and teacher alike. It is virtually impossible to teach very much statistics in a short period of time, nor is it practicable to devote a large portion of the measurement course to this area.

The writers have tried to solve this dilemma by omitting the more advanced material, by putting emphasis on concepts rather than computations, by repeating main ideas frequently, and by saving certain techniques for the Appendix. Fifty multiple-choice items are presented in Appendix A, pages 429–435, to help the student test his grasp of basic principles. A summary of common statistical terms and a selected list of statistics textbooks appear at the end of the chapter.

Despite these aids, the reader will find that he cannot read the chapter like a light novel. It will require careful study, much as one would prepare chemistry or physics assignments. However, this effort should result in permanently improved ability to understand some forms of statistical communication, though real mastery cannot be expected. Very likely your teacher will want to be quite selective in his emphasis upon the various topics in this chapter. He may skip some of them entirely.

B. Classification and Tabulation

Before test scores or other quantitative data can be comprehended and interpreted, it is usually necessary to summarize them. Table 3 gives a class record for a reading readiness test administered at the beginning of the school year. The scores appear in alphabetical order as they are recorded in the teacher's class roll book. However, the scores do not mean very much in this form. It is with some difficulty that we can tell whether Richard A, with a score of 90, for example, is a very superior or just an average pupil.

Rank Order. Ordinarily the first step is to arrange the scores in order of size, usually from high to low. This is called an *ungrouped series*. In a small class, this is sometimes all that is necessary. Table 4 gives the same scores as Table 3 arranged in order of size. This table also shows the *rank order* of the pupils and the scores tabulated without further grouping. It is now easy to see that Richard A's score of 90 gives him a rank of thirteen in a class of thirty-eight, or about one third of the way from the top. In a similar manner, it is easy to interpret each of the other scores in terms of rank. But this method, especially in classes of twenty or more pupils, is likely to prove unsatisfactory. Note, for example, that two pupils make a score of 97. Since it is not correct to say that one ranks higher than the other, it is necessary to assign them fractional ranks. As there are six pupils who rank higher, the next two ranks, 7 and 8, are averaged, which gives 7.5. In like manner the average of ranks 9 and 10 is 9.5, and so on for the other pupils with tied scores. Since there are three pupils each of whom makes a score of 75, and there are twenty-one pupils who rank above this score, 2

average of the next three ranks, 22, 23, and 24 is 23, which is the rank assigned each of the scores of 75. In addition to the fact that time and trouble are required to determine these ranks, the list is long and unwieldy to handle, and is inadequate for making comparisons with other classes which are much larger or much smaller.

The frequency table or distribution. One way out of the difficulty is to arrange the scores in a special way. Such a process is called *tabulation*. The table itself is called a *frequency distribution*, or merely *distribution*.

TABLE 3

A CLASS RECORD FOR A READING READINESS TEST
(38 Pupils)

<i>Pupil</i>	<i>Score</i>
Richard A	90
Robert B	66
Barbara B	106
Charles B	84
Mildred C	105
Robert C	83
Robbin C	104
Diney D	82
Jim D	97
John D	97
Robert D	59
Don F	95
Larry F	78
Richard G	70
Warren H	47
Sylvia H	95
Robert H	100
Grover H	69
Jack K	44
Clarence K	80
Jerome I	75
Mary M	75
Billy N	51
Nancy O	109
Carrie P	89
Ralph R	58
Billy S	59
William S	72
Gretta S	74
George S	75
Robert S	81
Jack S	71
Richard S	68
Mary S	112
Jean T	62
Richard W	91
Dolores W	93
Carl W	84

The third and fourth columns of Table 4 show the simplest form of a distribution. Such a distribution consists of two columns: the various scores are arranged in one column in order of size and opposite each score is recorded in the other column the number of times it occurs. Each entry in the second column is called a *frequency* abbreviated *f* and the total is represented by *N*.

It is usually desirable, however, to carry the process one step further.

TABLE 4

READING READINESS SCORES FROM TABLE 3 ARRANGED IN ORDER OF SIZE AND RANK ORDER AND TABULATED

Order of Size	Rank Order	Tabulated Without Further Grouping	
		Score	Frequency (<i>f</i>)
112	1	112	1
109	2	109	1
106	3	106	1
105	4	105	1
104	5	104	1
100	6	100	1
97	7 5	97	2
97	7 5	95	2
95	9 5	93	1
95	9 5	91	1
93	11	90	1
91	12	89	1
90	13	84	2
89	14	83	1
84	15 5	82	1
84	15 5	81	1
83	17	80	1
82	18	78	1
81	19	75	3
80	20	74	1
78	21	72	1
75	23	71	1
75	23	70	1
75	23	69	1
74	25	68	1
72	26	66	1
71	27	62	1
70	28	59	2
69	29	58	1
68	30	51	1
66	31	47	1
62	32	44	1
59	33 5		
59	33 5		
58	35		
51	36		
47	37		
44	38		
			<i>N</i> = 38

average of the next three ranks, 22, 23, and 24 is 23, which is the rank assigned each of the scores of 75. In addition to the fact that time and trouble are required to determine these ranks, the list is long and unwieldy to handle, and is inadequate for making comparisons with other classes which are much larger or much smaller.

The frequency table or distribution. One way out of the difficulty is to arrange the scores in a special way. Such a process is called *tabulation*. The table itself is called a *frequency distribution*, or merely *distribution*.

TABLE 3
A CLASS RECORD FOR A READING READINESS TEST
(38 Pupils)

<i>Pupil</i>	<i>Score</i>
Richard A	90
Robert B	66
Barbara B	106
Charles B	84
Mildred C	105
Robert C	83
Robbin C	104
Diney D	82
Jim D	97
John D	97
Robert D	59
Don F	95
Larry F	78
Richard G	70
Warren H	47
Sylvia H	95
Robert H	100
Grover H	69
Jack K	44
Clarence K	80
Jerome I	75
Mary M	75
Billy N	51
Nancy O	109
Carrie P	89
Ralph R	58
Billy S	59
William S	72
Gretta S	74
George S	75
Robert S	81
Jack S	71
Richard S	68
Mary S	112
Jean T	62
Richard W	91
Dolores W	93
Carl W	84

The third and fourth columns of Table 4 show the simplest form of a distribution. Such a distribution consists of two columns, the various scores are arranged in one column in order of size, and opposite each score is recorded in the other column the number of times it occurs. Each entry in the second column is called a *frequency*, abbreviated *f*, and the total is represented by *N*.

It is usually desirable, however, to carry the process one step further

TABLE 4

READING READINESS SCORES FROM TABLE 3 ARRANGED IN ORDER OF SIZE AND RANK ORDER AND TABULATED

Order of Size	Rank Order	Tabulated Without Further Grouping	
		Score	Frequency (<i>f</i>)
112	1	112	1
109	2	109	1
106	3	106	1
105	4	105	1
104	5	104	1
100	6	100	1
97	7	97	2
97	8	95	2
95	9	93	1
95	10	91	1
93	11	90	1
91	12	89	1
90	13	84	2
89	14	83	1
84	15	82	1
84	16	81	1
83	17	80	1
82	18	78	1
81	19	75	3
80	20	74	1
78	21	72	1
75	22	71	1
75	23	70	1
75	24	69	1
74	25	68	1
72	26	66	1
71	27	62	1
70	28	59	2
69	29	58	1
68	30	51	1
66	31	47	1
62	32	44	1
59	33		
59	34		
58	35		
51	36		
47	37		
44	38		

$N = 38$

As a rule, there is so wide a range of scores that it is economical to group them according to size, such as a group including all scores from 110 to 114, inclusive from 105 through 109, inclusive, and so on. Each group is called a class. The complete grouping arrangement is usually referred to as a *grouped frequency distribution*. While there is no absolutely fixed rule for the number of classes, it is usually advisable to make *not fewer than twelve classes nor more than about fifteen*. To have fewer than twelve classes is to run the risk of distorting the results, while more than fifteen classes produces a table that is inconvenient to handle.

Making the frequency table There are four steps in making the ordinary grouped frequency distribution. These are illustrated in Table 5, using the scores given in Table 4.

1 *Determine the range*, which is one more than the difference between the highest score and the lowest. Of these scores, the highest is 112 and the lowest is 44, which gives a range of $(112 - 44) + 1 = 69$.

2 *Select the class interval*, which is the size of the groups into which the scores are to be classified. To do this, divide the range by 12, which gives the largest group, or class interval, to be used, and by 15, which gives the smallest class interval to be used. In this case, $69 \div 12 = 5.75$, and $69 \div 15 = 4.6$. Since it is impractical to use any class interval except a whole number, the fractions are disregarded and the next highest whole number is taken. The class interval will, therefore, be either 6 or 5. Of these class intervals it is best to choose the one which is more convenient to use. Odd numbered intervals have whole-number midpoints when the class limits are fractional (end in .5), so usually they are to be preferred over even numbered intervals which have fractional midpoints. The midpoint of an odd numbered group like 110-114, wherein there are 5 scores, is 112, but if a class size of 6 were used, the score limits being, for example, 109-113, the midpoint of this even numbered group would be 110.5, which might result in more complex computations. Therefore, a class interval of 5 is somewhat preferable to 6 when the class limits are fractional.

3 *Determine the limits of the classes*. The table must, of course, be long enough to include the highest score and the lowest score. To facilitate tabulation start each class with a multiple of the class interval. If the lowest class starts with 40, which is a multiple of 5, it will accommodate the lowest score 44. Each succeeding whole-number class limit will be 5 points above the one just below it. The next class will start at 45, the next at 50, and so on until the highest score, 112, is included in the class 110-114.

4 *Make the tabulation*. A short vertical line (tally) is drawn for each score opposite the class in which it falls. To make a tabulation it is not necessary to have all the scores arranged in order, for this process may require more time than the tabulation itself. In the original alphabetical list the first score is 90. In the "tabulation" column opposite the class which begins with 90 a vertical line is drawn to indicate the score. The next

score is 66 This falls in the class which begins at 65, so a line is made there In the same way, a line is placed in the column opposite the appropriate class for each of the other scores To indicate the fifth score in each class a diagonal line is drawn across the other four This makes it easier to count the tallies in each class

The finished table omits the steps by which it was made In the simplest form of a frequency distribution only two columns occur, the first of which shows the various classes, usually arranged in descending order, and the

TABLE 5

AN ILLUSTRATION OF THE PROCESS OF MAKING A GROUPED FREQUENCY DISTRIBUTION

Original Scores (from Table 3)	Steps in Making the Distribution		
90	Step 1 Determining the range		
66	Highest Score	112	
106	Lowest Score	44	
84	Range = Difference + 1 = 68 + 1 = 69		
105	Step 2 Selecting the class interval		
83	69 - 12 = 58 largest class interval desirable		
104	69 - 15 = 46 smallest class interval desirable		
82	(5 chosen because of convenience in tabulation)		
97	Steps 3 and 4 Determining the limits of the classes and making the tabulation		
97			
59			
95			
78			
70			
47			
95	Whole Number Limits of Classes	Tabulation	Frequency (f)
100			
69	110-114	/	1
44	105-109	///	3
80	100-104	///	2
75	95-99	///	4
75	90-94	///	3
51	85-89	/	1
109	80-84	///	6
89	75-79	///	4
58	70-74	///	4
59	65-69	///	3
72	60-64	/	1
74	55-59	///	3
75	50-54	/	1
81	45-49	/	1
71	40-44	/	1
68			
112			N = 38
62			
91			
93			
84			

second of which shows the frequency or the number of scores in each class. When two or more schools or grades are to be compared, it is usually best to include all the data in the same table. In that case there will be a column for the classes into which the scores are grouped and one for each of the schools or grades being compared. Table 6 shows a frequency table which combines the record of six schools on a certain test. The number of grouping intervals varies from 9 for School F to 17 for Schools A and D.

TABLE 6

DISTRIBUTION OF READING READINESS SCORES FOR SIX SCHOOLS IN A CERTAIN CITY

Score	School A	School B	School C	School D	School E	School F	All Six Schools
120-124				1			1
115-119							
110-114			1				1
105-109			3		2	2	7
100-104		3	2	2	5	3	15
95-99		6	4	4	4	5	23
90-94	5	2	3	5	6	10	31
85-89	4	4	1	4	4	1	18
80-84	2	3	6	6	4	8	29
75-79	10	5	4	4	1	2	26
70-74	6	2	4	7	6	4	29
65-69	9	4	3		4		21
60-64	4	5	1	2	1		13
55-59	1		3		1		5
50-54	1		1				2
45-49	1		1				2
40-44			1	2	2		5
35-39	1	1					2
30-34		2					2
25-29		1					1
20-24							
15-19							
10-14	1						1
A	45	38	38	37	40	36	234

The form of the table. A few words may be said about the mechanical make-up of the table as it occurs in printed or typed form. Each table bears a number. Either Roman or Arabic numerals may be employed, but the latter seem to be increasingly favored. The table number may be centered above the title of the table, or it may be given at the beginning of the title. The table usually starts with two horizontal lines and ends with a single horizontal line. Another horizontal line separates the column headings from the body of the table, and other horizontal lines separate any summarizing measures which may be given under the table proper. Vertical lines may be used to separate the columns, but usually no lines are drawn along the

margins of the page. It is considered good form to avoid abbreviations in the table whenever possible, and to make the title and headings full enough to indicate clearly the contents of the table.

A two-way table, scattergram, or scatter diagram. Table 7 shows the chronological, educational, and mental ages of the 20 students in a

TABLE 7

THE CHRONOLOGICAL EDUCATIONAL AND MENTAL AGES OF THE 20 PUPILS IN AN EIGHTH GRADE CLASS

Pupil	Ages Expressed in Months		
	Chronological (CA)	Educational (EA)	Mental (MA)
A	150	188	208
B	147	186	218
C	155	185	201
D	160	183	185
E	141	183	165
F	160	182	191
G	154	181	185
H	164	180	193
I	165	179	181
J	157	176	165
K	167	176	187
L	157	176	176
M	161	176	180
N	157	175	166
O	158	171	197
P	161	173	154
Q	179	171	180
R	152	167	165
S	160	167	177
T	158	165	164

certain eighth-grade class. It is sometimes helpful to compare at the same time pupils' scores on two measures. A two-way table, called a scattergram or scatter diagram, makes this easier. Table 8 contains a two-way distribution of mental and educational ages from Table 7.

Mental ages, grouped into class intervals of six months, make up the column headings, educational ages, grouped into class intervals of two months, constitute the rows. For example, Pupil A, with an MA of 208 months and an EA of 188 months, falls in the third column from the right, or 204-209 class, and in the top row, or 188-189 class. In like manner, the horizontal position of each pupil in the distribution shows his MA, and the vertical position shows his EA. A tendency will be observed for the scores to arrange themselves in a diagonal pattern from lower left to upper right. This means that, in general, pupils who are low in MA are low in EA, and pupils who are high in MA are high in EA. However, a few exceptions

TABLE 3

A TWO-WAY DISTRIBUTION OF MENTAL AGE AND EDUCATIONAL AGE FOR AN EIGHTH GRADE CLASS

(r = 65)

		Mental Age (MA) in Months												EA Frequency
		150- 155	156- 161	162- 167	168- 173	174- 179	180- 185	186- 191	192- 197	198- 203	204- 209	210- 215	216- 221	
Educational Age (EA) in Months	188- 189										A			1
	186- 187												B	1
	184- 185									C				1
	182- 183			E			D	F						3
	180- 181						G		H					2
	178- 179						I							1
	176- 177			J		L	M	K						4
	174- 175			N					O					2
	172- 173	P												1
	170- 171						Q							1
	168- 169													0
	166- 167			R		S								2
	164- 165			T										1
MA Frequency		1	0	5	0	2	5	2	2	1	1	0	1	20

stand out. For example, Pupil P, who is lowest in MA, is in the fifth row from the bottom in EA.

When the identification of individual pupils in the scattergram is unimportant, the totals only are entered in the appropriate squares (the cells). For Table 8 all such entries would be 1's, though in Table 18 page 92 numbers in the cells run as high as 4.

C Some Elementary Notions Concerning Quantitative Data

Concepts versus computations Two of the most important concepts that apply to various kinds of test data are *variability* and *central tendency*. These abstract notions are useful in summarizing the main features of a bewildering mass of figures. There are several commonly used measures of variability and of central tendency. It is possible to be an adept computer of these measures without having a clear grasp of their meaning. Likewise, it is possible to understand the concepts reasonably well without being a competent computer, though a purely verbal knowledge of them is likely to be unsatisfying and inaccurate.

In order to read test manuals and other measurement literature effectively, one needs a real understanding of the concepts of variability and central tendency. As an analyzer of test scores and similar material one needs to acquire facility in computing a number of statistics. Comparatively little of this computational ability can be picked up in a short measurements course. Perhaps all but the rudiments of calculation are best taught in a statistics course or an integrated measurements statistics sequence.

Let us defer the computational routines for a few pages in an attempt to consider the concepts themselves. An imaginary informal visit with a rookie teacher may help illustrate one aspect of variability and central tendency in a concrete situation. If you become curious concerning some of the statistics reported look at pages 81-85 for computational procedures.

Joe's grading dilemma Joe Doe, a brand new English teacher, is faced with the problem of marking his first batch of English themes. He asked the 31 pupils in his tenth grade class to "write about 500 words on your favorite sport." Now the papers are before him: the evening is young and he wonders how to assign grades.

"Well," says Joe, "These papers differ in lots of ways. Lee Littlesay wrote less than half a page, while Eve Effervescent managed to use up seven pages. Furthermore the quality of handwriting varies greatly from one student to another, the girls being in general much better than the boys. Some of these individuals can't spell either while others can. Hmm if a pupil misspells the same word ten times does that count off the same as misspelling ten different words once each? Some youngsters have something to say but say it ungrammatically and others are uninteresting but

Twenty-six themes struck Joe as being "satisfactory," while he finally managed to classify 5 as "unsatisfactory." Thus 5 students out of 31, 16 per cent of the class, failed. The other 84 per cent passed.

Three categories. Do the "satisfactory" themes deserve grades of A, or B, or C? Joe doesn't know, since his two-category grading system fails to reveal any differences among the 26 students who earned check marks. Therefore he rescores these 26 papers, assigning a plus (+) to the "excellent" themes and a plus-minus (\pm) to the remaining ones. This gives him the three-category frequency distribution in Table 10: $\frac{1}{2}$ (39 per cent) of the students submitted "excellent" themes, $\frac{1}{4}$ (45 per cent) turned in "satisfactory" ones, and $\frac{1}{8}$ (16 per cent) were "unsatisfactory."

Four categories. It is easy enough for Joe to decide that the lowest 5 pupils deserve grades of "F" and the middle 14 should get "C's," but what about the top 12 persons? Not all of them did "A" work, thinks Joe, so he goes through the 12 best papers again, dividing them into 4 "A's" and 8 "B's." This gives him a new frequency distribution, the four-category one. Now the percentages are: A, 13 per cent, B, 26 per cent, C, 45 per cent, and F, 16 per cent. Thank goodness, muses Joe, I've finally spread these papers out enough to assign the usual grades. The simple check marks I started with didn't disperse the pupils at all, because 100 per cent of them got the same grade. Now less than half of the individuals earn any one grade.

Variability. Dimly Joe recalls something his measurement teacher said about "scatter," "dispersion," "heterogeneity," *variability*. All we do with letter grades of A, B, C, and F is to put each pupil into one of four ordered categories: A being the highest and F the lowest. That's what college "quality" or "honor" points were for, to indicate the rank or value of each category. A grade of A carried 3 quality points, B 2, C 1, and F 0. A bit of fancy calculation by Joe reveals that the range of the middle 50 per cent of the English class (that is, excluding the highest fourth and the lowest fourth) is from 0.70 to 2.03 quality points—a distance of 1.33 quality-point units. The type of computational procedure that he used is explained on pages 81-85.

Joe becomes curious about the range of the middle 50 per cent in the preceding distributions. Going back to the three-category distribution in Table 10 he calls the minus category 0, the plus-minus category 1, and the plus category 2. This makes the range of the middle half of the class 1.16 quality-point units, 0.17 less than when there were four categories.

For the two-category distribution, where a check mark is 1 and a minus 0, the middle half range is only 0.60, 0.56 less than the 1.16 secured for three categories and less than half the range of 1.33 for four categories. This pleases Joe, since it seems to confirm his hunch that variability increases in direct relationship to the number of categories. If this is so he concludes, why not put each of the 31 individuals into a category by himself—

that is, why not rank-order the papers from 1 (highest) to 31 (lowest), with no ties? Then the range of the middle 50 per cent would be 15.5

Eleven categories. The rank ordering goes along fairly well with the *A* and the *F* papers, but it is troublesome indeed for those who earned *C*'s. Despite his best efforts, Joe ends up with quite a few ties. The final result is that he settles for 11 categories rather than 31, which he labels *A+*, *A*, *A-*, *B+*, *B*, *B-*, *C+*, *C*, *C-*, *D*, and *F*, as shown in Table 10. The middle-50 percentile range of the scores in this frequency distribution is 3.5, much larger than the 1.33 range for four categories. This represents just about all the fineness of grading Joe feels he can obtain for an overall score on these themes.

Joe realizes that some English classes will vary more in the quality of themes submitted than will others. The college-preparatory class is probably less variable with respect to theme-writing ability than is his "general" section, so in that group the dispersion of grades would probably be less. This depends considerably on the grading philosophy of the individual teacher, of course, and on the care with which he goes over the themes. Some teachers give mainly *B*'s, while others give both *A+*'s and *F*'s. The teacher with the grading system most extreme in variation would be one who gives half the class *A+*'s and the other half *F*'s, with no *B*'s, *C*'s or *D*'s.

Variability of theme grades is dependent upon (1) real differences in ability among the students, (2) the nature of the theme and the theme-writing situation, and (3) the skill and care with which the themes are graded.

The vocabulary of variability. The middle-50 percentile range, 3.5, that Joe computed above is known technically as the *interquartile* range, the range between the boundaries of the upper and lower quarters of the group. More commonly, the *semi interquartile* range, which is half of the middle-50 percentile range ($3.5 - 2 = 1.75$), is reported. This is called *Q*, the *quartile deviation*, since it is a range within which about a quarter of the scores lie.

A better measure of variability or dispersion is the range of the middle 80 percentile of the scores, called *D*, but it has not been used as often as *Q*. For the scores in the 11-category distribution, $D = 6.9$.

The "crude" range, called simply the *range*, is the whole length of the distribution from the bottom of the lowest score to the top of the highest one. Start counting with 0 and go through 10, and you will see readily that the range of the 11-category distribution is 11. The range is a relatively unreliable measure of variability in most instances, so it is not used very often.

The best measure of variability for a large variety of situations is the *standard deviation* (abbreviated σ or SD). Twice the SD gives the range within which approximately the middle two-thirds of all the scores lie.

For the distribution of scores 0-10 in Table 10, it is 2.5. The greater the "range of talent" within a group of persons tested with a particular test, the greater the standard deviation of that group will be. A group for whom the standard deviation of scores is large is said to be "heterogeneous," while one in which all individuals earn about the same scores is "homogeneous." These are relative terms, however. Virtually all groups show some scatter on any test.

You would expect the geography test scores of pupils in Grades 4, 5 and 6 combined to be more variable than the scores of Grade 5 alone on the same test because probably a majority of the Grade 4 students know less geography than the average fifth grader, while most of the sixth graders know more. This would cause the standard deviation of the total group to be substantially larger than the SD of scores for any one of the three grades.

Administer a third grade spelling test to college sophomores and most of them will get all the words right, so variability of the scores will be slight. Give a seventh grade arithmetic test to third graders and most of them will miss almost everything. In each instance the standard deviation of the scores will be quite small compared with the SD that would have resulted had the test been given to a group for which it was of adequate difficulty.

The concept of central tendency. Because scores on a test vary from person to person, some statistic like the standard deviation is needed to summarize the extent of this variability without relying solely on intuitive study of the whole frequency distribution. Likewise, because the various persons tested earn different scores, it is not possible to cite any one figure that is completely typical of all individuals. The "modal" score or crude mode of the 11-category distribution in Table 10 is 3, since more persons (7) received it than any other single score.

The middle score category is 5, but 19 scores lie below it and only 9 lie above. The nearest thing to a really middle score is 4, with 15 frequencies below and 12 above.

Actually, the point that cuts the distribution into halves, with 15.5 frequencies falling above that point and 15.5 falling below, is 3.625. This is known as the *median*. It is the statistic with which Q , the quartile deviation or semi interquartile range, is usually linked. The range Q distance on each side of the median [$3.625 \pm (3.5 - 2) = 1.9$ to 5.4] is similar to the interquartile range. Here the interquartile range is 2.4 to 5.9 with a midpoint at 4.167, which is 0.54 higher than the median.

When your arithmetic text referred to the average, it meant the sum of all the scores divided by the total number of scores. This is often called the arithmetic mean, or simply the *mean*. Like the mode and the median, it is a measure of *central tendency*, indicating the point in the frequency distribution, usually near the center, toward which the scores tend. For the

11-category distribution the mean is 129 divided by 31, or 4.2. Compare this with the median of 3.6 and the crude mode of 3.

The mean grade is about $\frac{1}{5}$ th of the distance between C+ and B-. The median grade is $\frac{5}{8}$ ths of the distance between C and C+. The modal grade is C. The mode is based upon few cases and therefore too untrustworthy for use except as a rough, quick estimate of central tendency. The mean is the preferred measure, except in certain special cases.

Mean versus median. Suppose that the D category in the frequency distribution had been divided into three parts, with new low scores of -1 and -2 as shown in table 11.

TABLE 11
DIVIDING THE FOUR D'S FROM THE
11 CATEGORY DISTRIBUTION OF
TABLE 10 INTO THREE PARTS

Grade	Score	f
D+	1	2
D	0	1
D-	-1	1
F	-2	1

This does not change the value of the median or the mode, but the new mean is 124 divided by 31, which equals 4.0, a decrease of 0.2.

If a factory has 100 employees, all but five of whom earn between \$2000 and \$5000 per year, and if these five executives each earn more than \$10 000 the *mean* salary for the factory is likely to be misleadingly high. However, the median will not be sensitive to the few high salaries. In fact, the median can be ascertained without even knowing the actual salaries of the executives by just having a top category of "more than \$5000" whose frequency is 5. The use of the median instead of the mean may be desirable in such instances. Sometimes the better procedure is to exclude the five executives from the distribution and to report their salaries separately. Then the 95 workers represent a more homogeneous group for which the mean may be used.

An improbable anecdote should make this clear. Five men once sat together on a park bench. Two were vagrants, each with total worldly assets of 25 cents. The third was a workman whose bank account and other assets totaled \$2000. The fourth man had \$15 000 in various forms. The fifth was a multimillionaire with a net worth of \$5 000 000. Therefore the modal assets of the group were 25 cents. This figure describes two of the persons perfectly, but it is grossly inaccurate for the other three. The median figure of \$2000 does little justice to anyone except the workman. The mean of \$1 003 400 is not very satisfactory even for the multimil-

lonaire If we had to choose one measure of central tendency, very likely it would be the mode, which describes 40 per cent of this group accurately But if told that "the modal assets of five persons sitting on a park bench are 25 cents," we would be likely to conclude that the total assets of the group are approximately \$1 25, which is about \$5,000,000 lower than the correct figure Obviously, no measure of central tendency whatsoever is adequate for these "strange bedfellows," who simply do not "tend centrally"

D. Finding the Mode, the Median, and the Mean

Central tendency. Characteristic of most frequency distributions is a tendency for the scores to bunch or concentrate somewhere near the center An important statistic is, therefore, the point on the scale around which the scores tend to group themselves This is a measure of *central tendency* It is that value which typifies, or best represents, the whole distribution

One might wish to know which of several schools made the best record on a certain test, and which the poorest To determine this, compute an average for each school, and then note which one has the highest average and which one has the lowest average In other words, that school is best which *on the average* makes the highest score, and that school is poorest which *on the average* makes the poorest score²

Statisticians employ three common averages These are the mode, or inspectional average, the median, or counting average, and the mean, or computed average The meaning of each of these will now be considered

The mode. The most frequent score is called the *mode* It is obtained by inspection In Table 4 on page 63 the mode is 75, because more pupils made that score than any other The mode is not a very trustworthy average, however, especially with small groups In this case the changing of two scores might shift the mode decidedly If one of the pupils who made 75 had made 76, and if the one who made 58 had made 59, the mode would drop to 59, since more pupils would then have made that score than any other Largely because of its fickleness, the mode is not highly regarded as a measure of central tendency for small groups

The median. Perhaps the most widely used average in educational measurement is the *median* The median is the point which divides the distribution into halves Sometimes in an ungrouped series the *midscore* is used instead of the median Strictly speaking, when N is an even number, there is no midscore In that case, it is customary to average the middle pair of scores For example, in Table 4 on page 63 there are 38 pupils, 19 of whom made scores of 80 or less and 19 of whom made scores of 81 or more The midscore is then assumed to be the average of 80 and 81, the middle pair of scores, which equals 80.5 The terms *median* and *midscore* are often used

² Obviously, in order to be statistically and educationally significant, the difference between high and low schools should be fairly large

interchangeably, but the latter should be used only for scores arranged in order of size rather than in a grouped frequency distribution

TABLE 12
THE PROCESS OF LOCATING THE MEDIAN
(Scores from Table 5)

Frequency Distribution		Steps in the Process
110-114	1	Step 1 Obtaining $\frac{N}{2}$ $\frac{N}{2} = \frac{38}{2} = 19$
105-109	3	
100-104	2	
95-99	4	
90-94	3	
85-89	1	Step 2 Locating approximate median $1 + 1 + 1 + 3 + 1 + 3 + 4 + 4 = 18$ This takes us up to 79.5, which is the <i>approximate median</i>
80-84	6	
18		Step 3 Determining the correction $19 - 18 = 1$ $\frac{1}{6} \times 5 = \frac{5}{6} = 0.8$, the <i>correction</i>
75-79	4	
70-74	4	
65-69	3	
60-64	1	
55-59	3	
50-54	1	
45-49	1	Step 4 Locating the median. $79.5 + 0.8 = 80.3$ the median That is the median is the approximate median plus the correction
40-44	1	
N	38	

Table 12 illustrates the process of locating the median in a frequency distribution. The median is often described as the counting average, and it will be noted that counting does occupy a prominent place in its location. The steps may be summarized as follows:

1 Obtain $\frac{1}{2}N$. That is, divide the total of the frequencies by 2. Here $\frac{1}{2}N$ or $N/2 = 38/2 = 19$.

2 Locate the approximate median. Beginning at the low end of the distribution, count up the frequency column as far as possible without passing $N/2$ obtained in Step 1. In this case the frequencies $1 + 1 + 1 + 3 + 1 + 3 + 4 + 4$ give a total of 18. This is as far as we can go, for to include the next frequency, 6, would carry us too far, or beyond $N/2$, which is 19. The approximate median then is 79.5, halfway between the classes that include scores of 75-79 and 80-84.

3 Determine the correction needed. From $N/2$ subtract the total obtained

* Each class in the frequency table has fractional (.5) lower and upper limits, even though only whole number scores are shown. The class 75-79 actually runs from 74.5-79.5 and 80-84 means 79.5-84.5. If a score of 79.49 had existed in Table 3, it would have gone into the 75-79 class, while a score of 79.51 would have appeared in the 80-84 class. Also, the frequencies in the 80-84 class are considered to be evenly distributed over the 5-unit interval extending from 79.5-84.5.

in Step 2 In this case, $19 - 18 = 1$ This shows that one more score or unit is needed to obtain the required $\frac{1}{2}N$ scores And this score must come out of the next class, the 80-84 class, where there is a frequency of 6 That is, we must go $\frac{1}{6}$ of the distance into the next class As the class interval is 5, this means $\frac{1}{6}$ of 5, or 0.8 The correction is then 0.8

4 Obtain the median This is done by adding the correction to the approximate median In this case $79.5 + 0.8 = 80.3$, the median

$$5 \text{ Check by counting down } 84.5 - \left(\frac{19 - 14}{6} \times 5 \right) = 84.5 - \frac{5 \times 5}{6} = 84.5 - 4.2 = 80.3$$

The median for the eleven-category distribution in Table 10 on page 70 was secured by counting up in the following manner

1 Half of 31 is 15.5 $1 + 4 + 3 + 7 = 15$, which carries us up to 3.5

2 $15.5 - 15 = 0.5$, the number of frequencies to be gone up into the class that extends from 3.5-4.5 and has a total frequency of 4

$$3 \frac{0.5}{4} (1) = \frac{0.5}{4} = 0.1$$

4 $3.5 + 0.1 = 3.6$, the median

5 This checks with the result obtained by counting down

$$4.5 - \left(\frac{15.5 - 12}{4} \times 1 \right) = 4.5 - \frac{3.5}{4} = 4.5 - 0.9 = 3.6$$

The median is often used as a reference point for describing the location of individual pupils in a distribution A pupil in the higher half is said to be "above the median," and one in the lower half is said to be "below the median" Other points in the distribution are used in a similar manner For example, *quartiles* divide the distribution into fourths, and *deciles* divide it into tenths A pupil in the highest fourth is said to be "above Q_3 ," and one in the lowest fourth is said to be "below Q_1 ."

Quartiles should not be confused with *quarters* (fourths) of the distribution Persons scoring above Q_3 are in the highest fourth of the group but *not* in the "highest quartile," since this expression is meaningless Likewise, pupils scoring below Q_1 are in the lowest fourth but *not* in the "lowest quartile" There are only three quartiles, all of which are points rather than ranges Q_2 is the median, Q_0 and Q_4 do not exist

Similarly, there are 9 deciles, going from 1 through 9 A person may score at the 2nd decile (the 20th percentile) or between the 2nd and 3rd deciles, but *not* in the 2nd decile Rather, we would say that he scored in the third tenth of the group, counting from the bottom There are no 0th and 10th deciles

* Table 14 on page 82 illustrates the computation of the quartiles

The position of a certain pupil may be still more accurately described by indicating the percentage of pupils who fall below him. The points that divide a distribution into 100 equal divisions, or per cents, are called *percentiles* or, more simply, *centiles*.

Computation of percentiles. The median is the 50th percentile, since 50 per cent of all the frequencies lie below that point and 50 per cent lie above it. A percentile is a point in the score distribution below which the stated percentage of all measures lies. Thus an individual who scores at the 30th percentile of his class has done better than 30 per cent of the students and poorer than 70 per cent. Percentiles are computed in much the same manner as the median, the only difference being that the number of frequencies to be counted up depends upon the percentile desired.

The 30th percentile of the distribution in Table 12 is obtained as follows

1 30% of $N = 30 \times 38 = 11.4$ $1 + 1 + 1 + 3 + 1 + 3 = 10$, which carries us up to 69.5

2 $11.4 - 10 = 1.4$, the number of frequencies to be gone up into the class that extends from 69.5-74.5 and has a total frequency of 4

$$3 \quad \frac{1.4}{4} \times 5 = \frac{1.4 \times 5}{4} = \frac{7.0}{4} = 1.75$$

4 $69.5 + 1.75 = 71.25$, the 30th percentile

5 Check by counting down $100\% - 30\% = 70\%$ of the way

$$1 \quad 0.70 \times 38 = 26.6 \quad 1 + 3 + 2 + 4 + 3 + 1 + 6 + 4 = 24$$

$$2 \quad 74.5 - \left(\frac{26.6 - 24}{4} \times 5 \right) = 74.5 - \frac{2.6 \times 5}{4} = 74.5 - \frac{13}{4}$$

$$= 74.5 - 3.25 = 71.25$$

The 30th percentile in this distribution is 71.25, a point above which 26.6 (70 per cent) of the scores lie and below which 11.4 (30 per cent) fall.

At first it may be difficult for you to think of the 50th percentile as "average" because of the minimum passing percentage mark of 70 or 75 that is quite common in high schools. Of course, there is no such thing as a "failing" percentile. The decision to pass or fail a given student is an arbitrary one. For roughly descriptive purposes, however, the 25th and 75th percentiles are sometimes used as boundary lines, scores in the lowest fourth of a group are said to be "low," while those in the highest fourth are called "high."

The mean. The most familiar average is the *mean*, often called the *arithmetic mean*. In fact, this measure is in such common use that the ordinary person regards it as *the* average, because it is the only average he knows anything about. When the term "average" is met with in ordinary

conversation or the newspaper in such statements as "average temperature," "average rainfall," "average yield of corn and wheat," "average price," and the like, it is almost certain to be the mean that is meant. The mean can be computed merely by obtaining the sum of the measures and dividing by their number. The measure so obtained is then the value that each individual would have if all shared equally.

When the scores are few in number or in an ungrouped series, the simplest process of computing the mean is the one described above, that is, the scores are first added and then this sum is divided by the number of scores. For example, the sum of the 38 scores in Table 3 on page 62 is 3,050, and $3,050 \div 38 = 80.3$. When the scores are sufficiently numerous to justify the use of a frequency distribution, the so-called "short" method of computing the mean may be more convenient.

$$M = M' + \frac{i \times \Sigma fd}{N}$$

In this formula

M = mean,

M' = assumed mean (the midpoint of any class),

Σfd = sum of frequencies multiplied by their respective deviations (Σ indicates "sum of"),

N = total number of frequencies,

i = class interval

The method of computing the mean by the short formula is illustrated in Table 13. The steps in the process are as follows:

Step 1 *Assume a mean.* This is taken at the midpoint of some class. Any class may be selected, even though completely outside the frequency distribution, but choosing one near the center of the distribution makes the figures to be handled much smaller. In this case, the assumed mean is taken at the midpoint of the 80-84 class,

$$\text{which is } \frac{80 + 84}{2} = 82$$

Step 2 *Lay off the deviations from the assumed mean.* The plus deviations indicate how many classes various frequencies are above the assumed mean, and minus deviations indicate how many classes various frequencies are below the assumed mean. This column is headed d .

Step 3 *Multiply each f by its corresponding d .* This column is headed $f \times d$. The first product is $1 \times 6 = 6$, the second is $3 \times 5 = 15$, and so on.

Step 4 *Obtain the algebraic sum of the $f \times d$ column.* Note that the sum of the $+$ values is 48 and the sum of the $-$ values is -61 . The algebraic sum of -61 and 48 is -13 . Had the $+$ values exceeded the $-$ values, the sum would have been $+$. This is called Σfd .

TABLE 13
A SHORT WAY TO COMPUTE THE MEAN

Computation				Steps in the Process	
	<i>f</i>	<i>d</i>	<i>f</i> × <i>d</i>	Step 1 Assuming a mean. 82 is taken as the assumed mean. Two parallel lines indicate the class where it is the midpoint	
110-114	1	+6	+6	Step 2 Laying off deviations from assumed mean. This is the column headed <i>d</i>	
105-109	3	+5	+15		
100-104	2	+4	+8		
95-99	4	+3	+12		
90-94	3	+2	+6		
85-89	1	+1	+1	Step 3 Multiplying each <i>f</i> by its <i>d</i> . This column is headed <i>f</i> × <i>d</i>	
80-84	6	0			
75-79	4	-1	-4	Step 4 Obtaining algebraic sum of the <i>f</i> × <i>d</i> column. Sum of + values is 48. Sum of - values is -61. Algebraic sum is -13. This is Σfd .	
70-74	4	-2	-8		
65-69	3	-3	-9		
60-64	1	-4	-4		
55-59	3	-5	-15		
50-54	1	-6	-6	Step 5 Substituting proper values in the formula, as indicated at the lower left.	
45-49	1	-7	-7		
40-44	1	-8	-8		
$\Lambda = 38$			+48		
			-61		
			$\Sigma fd = -13$		
$M = M' + \frac{1 \times \Sigma fd}{\Lambda}$					
$M = \frac{80 + 84}{2} + \frac{5 \times -13}{38}$					
$= 82 + \frac{-65}{38} = 82 - \frac{65}{38}$					
$= 82 - 1.7 = 80.3$					

Step 5 Substitute in the formula $M = M' + \frac{1 \times \Sigma fd}{N}$. M' was found in Step 1 to be 82. The class interval, 1, is 5, since there are five whole numbers in each class. 40-44 means 40, 41, 42, 43, and 44. Σfd , found in Step 4, equals -13. $\Lambda = 38$, the total number of scores. Therefore

$$M = 82 + \frac{5 \times -13}{38} = 82 + \frac{-65}{38} = 82 - 1.7 = 80.3$$

It will be recalled that the mean when computed by the "long" method was 80.3 exactly the same as when computed by the "short" method. In other similar problems a discrepancy may occur due to the fact that the former method is based upon the actual value of the scores, whereas the latter method is based on the assumption that the midpoint of each class

is the average for all the scores in that class, an assumption which is often only approximately true. The difference in the result is usually so slight as to be of no practical importance.

What average is best? As a rule, the mean is regarded as the best measure of central tendency and the mode the poorest. The mean, however, is greatly influenced by extreme scores, and whenever it is desired to avoid this influence, the median is to be preferred. As such situations often arise in educational measurement, the median is widely used. For example, if the test is too difficult, there may be several zero scores, and if the test is too easy, there may be several perfect scores. But in neither case are the pupils at the extremes correctly measured. The median in some such situations is the best average to use. The median is also easier to find than the mean, unless an electric calculator is available.

E. Measures of Variability or Scatter

Meaning of variability. No distribution is completely described by its average or central tendency. Two classes in a school might have the same average intelligence and yet be very unlike. The members of one class may vary all the way from feeble-mindedness to the genius level, while all the members of the other group may rate as normal. Obviously, these two classes present different instructional problems because they differ in *variability*. Variability is the extent to which the scores tend to scatter or spread above and below the average. It is clearly important to have some convenient method of indicating the variability of a group. There are three common measures of variability: the range, the quartile deviation, and the standard deviation. All these measures represent distances rather than points, and the larger they are the greater the variability or scatter of the scores.

The range. The range has already been referred to as the distance between the lowest and the highest scores plus one.⁵ It is usually a very untrustworthy measure of variability.⁶ The shift in a single score may greatly alter the range and thereby materially increase or reduce the apparent variability of the group. School A and School D in Table 6 on page 66 illustrate this possibility.

The quartile deviation. A measure of variability that avoids being unduly influenced by extreme scores is the *quartile deviation*, or *Q*. This is one half the distance between the first and third quartiles. For this reason, it is often referred to as the semi interquartile range. Since 25 per cent of the scores fall below the first quartile, or Q_1 , and 25 per cent of the scores exceed the third quartile, or Q_3 , the interquartile range is the range of the

⁵ More precisely, the range is the difference between the lower real limit of the lowest class and the higher real limit of the highest class.

⁶ When there are only two measures ($N = 2$) however, the range gives all the information concerning variability that the distribution can yield.

middle 50 per cent of the scores. The whole interquartile range might be used to express the variability of the group, but it is customary to take only half this distance and to set up a new middle-half range extending from Q below the median to Q above the median $Mdn \pm Q$. As already noted, the middle of the interquartile range will not usually be the median, while the middle of this new range will always be. On the other hand, exactly half of all the frequencies lie within the interquartile range and exactly half outside it, but this does not hold precisely for $Mdn \pm Q$.

The formula used for obtaining Q is

$$Q = \frac{Q_2 - Q_1}{2} = \frac{75\text{th percentile} - 25\text{th percentile}}{2}$$

TABLE 14

THE PROCESS OF COMPUTING THE QUARTILE DEVIATION, Q

Frequency Distribution		Steps in the Process
110-114	1	Step 1 Computing Q_1 , the 25th percentile $\frac{1}{4}N = \frac{1}{4}$ of 38 = 9.5 Counting up 1 + 1 + 1 + 3 + 1 = 7, approximate Q_1 is 64.5 $9.5 - 7 = 2.5$, $\frac{2.5}{3} \times 5 = \frac{12.5}{3} = 4.17$, correction $64.5 + 4.17 = 68.67$, Q_1
105-109	3	
100-104	2	
95-99	4	
	28	
90-94	3	Step 2 Computing Q_3 , the 75th percentile $\frac{3}{4}N = \frac{3}{4}$ of 38 = 28.5 Counting up 1 + 1 + 1 + 3 + 1 + 3 + 4 + 4 + 6 + 1 + 3 = 28, Approximate Q_3 is 94.5 $28.5 - 28 = 0.5$, $\frac{0.5}{4} \times 5 = \frac{2.5}{4} = 0.62$ $94.5 + 0.62 = 95.12$, Q_3
85-89	1	
80-84	6	
75-79	4	
70-74	4	
65-69	3	
	7	Step 3 Substituting in formula. Formula $Q = \frac{Q_3 - Q_1}{2}$ Substituting $Q = \frac{95.12 - 68.67}{2} = 13.2$
60-64	1	
55-59	3	
50-54	1	
45-49	1	
40-44	1	
N	38	

Table 14 illustrates the computation of Q . It will be observed that the process of locating quartiles is like that of locating any other percentile. In the first step, the fractional part of N indicates the proportion of the distribution which falls below the desired point, that is, for Q_1 it is $\frac{1}{4}N$ and for Q_3 it is $\frac{3}{4}N$. There are four steps, as follows:

- Step 1. Compute Q_1 , the 25th percentile To begin with, $\frac{1}{4}$ of 38 is 9.5 The next three steps in locating this point are exactly the same as those in locating any percentile
- Step 2 Compute Q_3 , the 75th percentile Here the first step is to take $\frac{3}{4}N$, $\frac{3}{4}$ of 38 is 28.5 The other three steps are identical with those in locating any percentile
- Step 3 Substitute in the formula Q_2 is 95.12 and Q_1 is 68.67 The difference between them is 26.45 Half of this difference is 13.2, the value of Q
- Step 4 Check your work by counting downward

The interpretation of Q and other measures of variability is a relative matter Whether a Q of 13.2 is to be considered great or small depends upon the magnitude of comparable measures for other groups using the same test

The standard deviation. A third measure of variability, which has many uses in educational measurement is the *standard deviation*. It is usually represented by the Greek letter σ , called sigma, and defined as the square root of the mean of the squares of the deviations of the scores from their mean. It may also be defined as that range above and below the mean ($M \pm 1\sigma$) that in a normal distribution⁷ includes 68.26 per cent, or approximately two thirds, of the scores

The formula for the standard deviation when computed from an assumed mean for scores in a frequency distribution is

$$\sigma = \frac{1}{N} \sqrt{(N \times \sum fd^2) - (\sum fd \times \sum fd)} = \frac{\sqrt{N \sum fd^2 - (\sum fd)^2}}{N}$$

The computational process is illustrated in Table 15. It can be seen that the only term not used in the computation of the mean in Table 13 is $\sum fd^2$, the sum of each frequency times the square of its respective deviation. The steps needed for computing the standard deviation are as follows:

- Step 1 Assume that the mean falls at the midpoint of a certain class, say 80-84
- Step 2 Lay off the deviations above and below the assumed mean
- Step 3 Multiply each f by its d
- Step 4 Obtain $\sum fd$, the algebraic sum of the fd column. Here it is -13
- Step 5 Prepare the $d \times (f \times d)$ column. Each entry in this column is the product of a d and the $f \times d$ to its right
- Step 6 Obtain $\sum fd^2$, the sum of the $d \times (f \times d)$ column. All values in the $d \times (f \times d)$ column are positive, since negative deviations are squared. Their sum is 479
- Step 7 Substitute in the formula for σ

⁷ This particular type of frequency distribution is discussed on pages 260 and 290

TABLE 15

A SIMPLIFIED WAY TO COMPUTE THE STANDARD DEVIATION

Computation					Steps in the Process
	<i>f</i>	<i>d</i>	<i>f</i> × <i>d</i>	<i>d</i> × (<i>f</i> × <i>d</i>)	
110-114	1	+6	+6	36	Step 1 Assume a class for the mean (80-84)
105-109	3	+5	+15	75	Step 2 Lay off deviations from assumed mean
100-104	2	+4	+8	32	Step 3 Multiply each <i>f</i> by its <i>d</i>
95-99	4	+3	+12	36	Step 4 Obtain Σfd the algebraic sum of the <i>f</i> × <i>d</i> column Here it is -13
90-94	3	+2	+6	12	Step 5 Prepare the <i>d</i> × (<i>f</i> × <i>d</i>) column Each entry is the product of <i>d</i> and the <i>f</i> × <i>d</i> opposite it
85-89	1	+1	+1	1	Step 6 Obtain Σfd^2 This is merely the sum of the <i>d</i> × (<i>f</i> × <i>d</i>) column Here it is 479
80-84	6	0			Step 7 Substitute in the formula as indicated.
75-79	4	-1	-4	4	
70-74	4	-2	-8	16	
65-69	3	-3	-9	27	
60-64	1	-4	-4	13	
55-59	3	-5	-15	75	
50-54	1	-6	-6	36	
45-49	1	-7	-7	49	
40-44	1	-8	-8	64	

$$N = 38 \qquad \qquad 48 \qquad 479$$

$$\qquad \qquad \qquad -61 \qquad \Sigma fd^2$$

$$\Sigma fd = -13$$

$$\sigma = \frac{\sqrt{N \Sigma fd^2 - (\Sigma fd)^2}}{N} = \frac{5\sqrt{38(479) - (-13)^2}}{38} = \frac{5\sqrt{18202 - 169}}{38} =$$

$$\frac{5\sqrt{18033}}{38} = \frac{5 \times 134.3}{38} = \frac{671.5}{38} = 17.7$$

D, a useful percentile measure of variability. σ may be estimated fairly accurately and easily by means of the formula

$$\sigma = 0.4 \times D = 0.4 \times (90\text{th percentile} - 10\text{th percentile})$$

For the scores in Table 15, $0.4D$ would be secured as follows

Step 1 Obtain the 90th percentile, which is a point that lies 10 per cent of the way down into the score distribution. 10 per cent of $38 = 3.8$, so the 90th percentile = $109.5 - \left(\frac{3.8 - 1}{3} \times 5\right) = 109.5 - \frac{2.8 \times 5}{3}$

$$= 109.5 - \frac{14.0}{3} = 109.5 - 4.67 = 104.83$$

Step 2 Obtain the 10th percentile, which is a point that lies 10 per cent of the way up into the score distribution. The 10th percentile = $51.5 + \left[\frac{3.8 - (1 + 1 + 1)}{3} \times 5\right] = 51.5 + \frac{0.8 \times 5}{3} = 51.5 + \frac{4.0}{3}$

$$= 51.5 + 1.33 = 52.83$$

Step 3 Substitute in the formula $\sigma = 0.4D$ $\sigma = 0.4 \times (101.83 - 55.83) = 0.4 \times 49 = 19.6 = 20$ Contrast this with the 17.7 value of σ in Table 15, which rounds off to 18

Though Q is used much more frequently than D , the latter is a far better percentile measure of variability and is considerably easier to compute

Practical uses of the standard deviation. The standard deviation is the most important measure of the variability of test scores. A small standard deviation means that the group has small variability or is relatively homogeneous, while a large standard deviation means the opposite condition, heterogeneity. σ also has certain other important uses.

The position of a pupil in a distribution is often represented in terms of standard deviation units. In the distribution used in Tables 13 and 15, where the mean is 80 and the standard deviation is 18, a pupil whose score is 98 is said to be one standard deviation above the mean, and the score is written $+1\sigma$. In like manner, a pupil whose score is 62 is said to be approximately one standard deviation below the mean, and the score is written -1σ . Such scores are called *standard scores* or *z scores*.⁸

Which measure of variability is best? As a rule, σ is regarded as the best measure of variability, and the range is undoubtedly the poorest. The range is subject to all the limitations which the mode has as a measure of central tendency. Just as the mean is greatly influenced by extreme scores so is σ . Whenever it is desirable, therefore, to avoid the influence of extreme scores, the median is employed as a measure of central tendency, and with it a percentile measure of variability such as Q or D . In like manner, when the mean is used, σ is the appropriate measure of variability.

F. Measures of Relationship

The concept of co-relationship or concomitant variation. During the latter part of the nineteenth century, Sir Francis Galton and the pioneer statistician Karl Pearson succeeded in developing the theory and mathematical basis for what is now known as *correlation*.⁹ They were concerned with relationships between two variates, for example, height and weight. It is easy to note that tall persons usually weigh more than short ones, suggesting that above-average height tends to go with above average weight. Height and weight vary together though certainly not perfectly, there are "beanpoles" and "five by fives" to upset the relationship. It would be possible to select a group of individuals in such a manner that the taller the person is the less he weighs, but this negative relationship between height and weight is not to be expected for individuals picked at random.

⁸ For a fuller discussion of *z scores* see page 289.

⁹ A fascinating account is contained in Helen M. Walker *Studies in the History of Statistical Method*, Chapter V. Baltimore: The Williams & Wilkins Company, 1929. The concepts of correlation and regression are treated together in most statistics texts but for the sake of simplicity regression is not mentioned in the present discussion.

Let us examine some other factors which normally vary together. There is a substantial, but again by no means perfect, positive correlation between intelligence test scores and average grades earned during the freshman year of college. The higher the score obtained by the entering freshman, the higher his grades are likely to be. The lower an individual's score, the poorer a student he will probably make. This relationship has been found with all sorts of intelligence tests used in a great number of different type colleges ever since such tests first became available commercially shortly after the close of World War I.¹⁰

Husbands and wives tend to be more like each other with respect to age, amount of education, and many other factors, than they are like people in general. The sons of tall fathers tend to be taller than average, and the sons of short fathers tend to be short. Likewise, the fathers of tall sons tend to be of above average height. Children resemble their own parents in intelligence more closely than they resemble other adults. Positive correlation between members of families is usually found for almost any characteristic from algebra ability to knowledge of zoology.

Using Galton's ideas concerning trait resemblances, Pearson devised as a measure of relationship the *product-moment coefficient of correlation*, r . Since about 1900 this has been a widely employed statistic. In the testing field it has become almost indispensable. Virtually all test manuals are plentifully sprinkled with r 's, as is most educational literature. The classroom teacher frequently encounters r 's in his reading and conversation. For these reasons, both the concept and the computational procedure are well worth mastering.

Pearson's original r and several other related r 's summarize the magnitude and direction of the relationship between two sets of measurements, such as height and weight based upon the *same* persons, or between measurements on *pairs* of persons, like the fathers and sons mentioned above. It makes no difference whether the variates are history grades and geography grades, or speed of running the hundred-yard dash and skill in playing the violin, or speed of tapping and age. In every situation, r can have values that range from -1 for a perfect inverse relationship through 0 for no systematic correlation to $+1$ for perfect direct relationship, and the r 's between radically different kinds of variates are wholly comparable. For example, it is meaningful to say that reading ability and intelligence are more closely related than height and weight.

In order to compute r without an electric calculator, it is helpful to have a scatter diagram such as Table 8 on page 68. However, those scores are too simple to illustrate computational procedures well, since not more than

¹⁰ One of the most complete summaries is Harley F. Garrett, "A Review and Interpretation of Investigations of Factors Related to Scholastic Success in Colleges of Arts and Sciences and Teachers Colleges," *Journal of Experimental Education*, 18: 91-118, December 1919.

a single individual falls into any one cell of the scattergram. The r between MA and EA for Table 8 is .65, which indicates a moderate positive relationship.

TABLE 16

A SCATTER DIAGRAM ILLUSTRATING NEGATIVE CORRELATION ($r = -.35$) BETWEEN CHRONOLOGICAL AGE AND EDUCATIONAL AGE FOR 20 EIGHTH GRADERS
(Data from Table 7)

		Chronological Age (CA) in Months													EA Frequency
		141-143	144-146	147-149	150-152	153-155	156-158	159-161	162-164	165-167	168-170	171-173	174-176	177-179	
Educational Age (EA) in Months	188-189				1										1
	186-187			1											1
	184-185					1									1
	182-183	1						2							3
	180-181					1			1						2
	178-179									1					1
	176-177						2	1		1					4
	174-175						2								2
	172-173							1							1
	170-171													1	1
	168-169														2
	166-167				1			1							1
	164-165						1								1
CA Frequency		1		1	2	2	5	5	1	2				1	20

Negative correlation is illustrated in Table 16, using the 20 chronological-age and educational-age scores from Table 7, page 67. Since the older children in a school grade are usually less able students than the younger ones, pupils with high CA's tend to have low EA's, and those with low CA's tend to have high EA's. That this relationship in Table 16 is rather weak is indicated by the low magnitude of the r , which is $-.35$. The oldest pupil has a CA of 179 and an EA of 171, while the youngest pupil's CA is 141 and his EA 183. However, one of the younger pupils secured an EA of only 167.

Two urgent cautions are imperative. First, *r cannot be interpreted directly as a percentage*. An r of 0 represents no relationship at all, but an r of .65 does *not* mean 65 per cent relationship. In one sense, the difference in degree of relationship represented by r 's of .91 and .98 is as great as the difference between r 's of 0 and .65. As r 's get larger, a small gain indicates a considerable increase in the degree of correlation. Therefore, an r of .66 indicates *more* than twice the relationship shown by an r of .33. These relative magnitudes are depicted graphically in Figure 4¹¹ where the curve for negative r 's is symmetrical with the one for positive r 's. Thus a given r indicates a high degree of relationship if its magnitude, regardless of sign, is large. An r of $-.72$ denotes just as strong an inverse relationship as an r of .72 indicates direct covariation, both have equal predictive value. Note that the two curves of Figure 4 are approximately linear through about the first fourth of the r scale, after which they become increasingly positively accelerated. An r of .24 corresponds to a z of .24, while an r of .995 corresponds to a z of 3.00.

The second warning is that *correlation does not necessarily mean causation*. Often variables other than the two under consideration are responsible for the association. Furthermore, problems in the social sciences, the field in which correlation is most used, are usually too complex to be explained in terms of a single cause.

Let us take several examples. It is probably true that in the United States there is moderate positive correlation between the average salaries of teachers in various high schools and the percentages of their graduates who go on to college, but to say that these students attend college *because* their teachers are well paid is as inaccurate as to say that their teachers are well paid because many of the graduates attend college. The situation is complex, but one prominent factor is the financial condition of the community, which to a considerable extent determines ability to pay *both* teachers' salaries and college expenses.

Furthermore, it has been found that the percentage of "dropouts" occurring in high schools varies inversely with the number of books per pupil in the libraries of those schools.¹² But common sense tells us that piling more books into the library will hardly affect the dropout rate, nor will getting a better attendance officer bring about a magical increase in the number of books.

Failure to recognize the non-causal nature of correlation is, in its broadest sense, a widespread logical error, for the fundamental notions of co-rela-

¹¹ Based upon Ronald A. Fisher, *Statistical Methods for Research Workers* (Tenth Edition) Table V B, page 210. London: Oliver and Boyd, 1948. The vertical (ordinate) figures, ranging from 0 to 3.0, are Fisher's z transformation values of the r scale. They are not the same as the z -scores mentioned on page 85.

¹² For several such relationships, see Guy V. Ferrell, *High School Holding Lower—An Analysis of Certain Internal Factors*. Unpublished Ph.D. Dissertation, George Peabody College for Teachers, 1951, 228 pages.

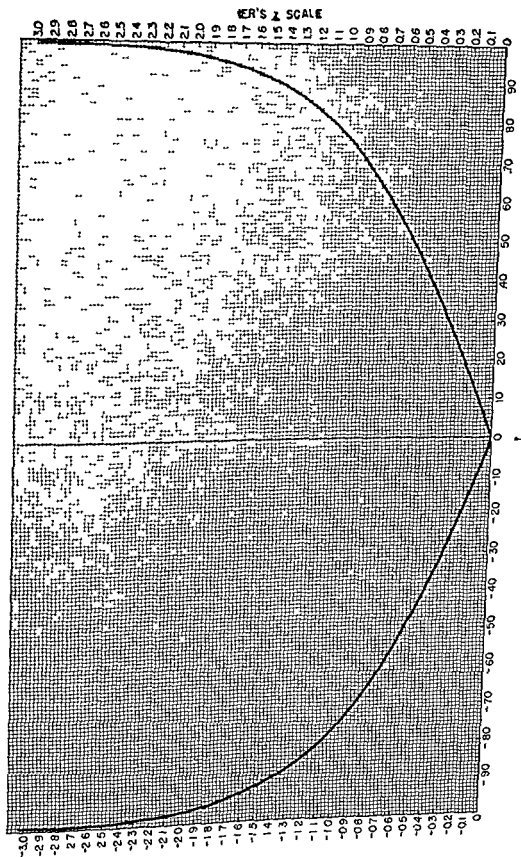


Figure 4 The Relative Amount of Relationship Represented by r_s of Various Sizes

tionship affect our lives at many points. Going to Sunday School is generally agreed to be valuable from many standpoints, but a positive relationship between the rate of attendance and a characteristic such as honesty does not *necessarily* imply that children are honest because they attend Sunday School. Underlying both attendance and honesty may be "home training," for example. A really crucial test of the hypothesis that attending Sunday School "makes" children more honest would have to be experimental rather than correlational.

Note carefully that while correlation does not directly establish a "causal" relationship, it may furnish *clues* to causes—and these can be taken advantage of in planning controlled experimentation. Therefore, r is useful primarily for exploratory purposes. Hence, it is employed much more widely in the newer sciences such as sociology, psychology, and education than in physics and chemistry.

Various ways to obtain r . There are many devices for computing r . One of the simplest methods requires an electric calculator and utilizes the "raw" scores themselves, without any frequency distributions or scattergrams. However, it is seldom feasible unless a calculator is available, because the arithmetical operations involve large numbers and therefore become quite tedious. Almost all other computational procedures require a scattergram. Many routinized "correlation charts" are available commercially, and nearly every statistics teacher has his own pet version, usually to some extent original.

This diversity of computing aids is probably indicative of basic difficulties inherent in the process of computing r "by hand." There is not any really simple way to do it. The chief difficulty is that r indicates a relationship between paired scores, so in order to obtain r without undue labor one must by some *indirect* method find the sum of the products of the paired scores. That is the intent of every simplified procedure.

One complicating factor in the attempt to simplify the computation of r is that in order to make the numbers involved as small as possible, most chart designers set up procedures that result in many negative numbers. Negative numbers are likely to confuse the average student, while large positive ones are tedious to handle. Both invite sizable errors. In the writers' opinion, it is better for persons who are getting their introduction to r in this book to work with somewhat larger positive numbers rather than with smaller negative ones, since, thereby, the explanation is simplified considerably and the chances of misunderstanding the procedure reduced. If you already know how to compute r from a scattergram by some other method and know it well, use it rather than the one described below.¹¹

Constructing the scattergram. In order to get some "live" data, one of the authors took 43 sets of classroom test scores from his roll book.

¹¹ A thorough explanation of one procedure is given by Helen M. Walker, *Elementary Statistical Methods*, pages 229-233. New York: Henry Holt and Company, 1913.

These are shown in Table 17. Note there that the X scores represent number of points (not percentage right) on a pretest at the beginning of the quarter, while the Y scores were secured at midterm.

TABLE 17

PRETEST AND MIDTERM SCORES OF 43 GRADUATE STUDENTS ON TWO TEACHER MADE OBJECTIVE TESTS IN INTERMEDIATE STATISTICS

Student	Pretest Score (X)	Midterm Examination Score (Y)	Student	Pretest Score (X)	Midterm Examination Score (Y)
a	62	51	v	62	65
b	55	66	w	53	56
c	55	40	x	62	56
d	49	38	y	49	54
e	46	51	z	62	56
f	67	57	aa	44	39
g	32	42	bb	60	49
h	42	35	cc	47	55
i	67	61	dd	44	45
j	55	46	ee	49	39
k	44	33	ff	36	38
l	46	58	gg	40	15
m	37	48	hh	53	50
n	57	41	ii	44	47
o	34	55	jj	49	41
p	57	44	kk	49	35
q	58	59	ll	44	43
r	58	45	mm	53	60
s	49	51	nn	53	52
t	40	32	oo	42	35
u	73	54	pp	57	45
			qq	49	41

The highest score in the X distribution is 73 and the lowest 32, so the range is $(73 - 32) + 1$, or 42. Since $\frac{42}{4} = 10.5$, and $\frac{42}{3} = 14$, it would be appropriate to use a grouping interval of either 4 or 3. Trying 4 and starting the whole-number lower limit of the lowest class with 32, which is a multiple of 4, we find that the classes run from 32-35 to 72-75 and that there are only 11 classes. Since our arbitrary rule is to have not fewer than 12 nor more than approximately 15 classes, we need to use the smaller interval, 3. With it there will be 15 classes running from 30-32 (30 is a multiple of 3) to 72-74.

Repeating this kind of procedure for the Y distribution, where the scores range from 66 to 15, results in a grouping interval of 4. The 14 classes run from 12-15 (12 is a multiple of 4) to 64-67.

Now construct a 15×14 scattergram, as shown in Table 18.¹⁴ Put the

¹⁴ The method set forth in Table 18 is a slight variation of one described in Quinn McNemar, *Psychological Statistics* pages 94-96. New York: John Wiley & Sons, 1949.

TABLE 18
COMPARISON OF THE r BETWEEN SCORES ON TWO TEACHER-MADE OBJECTIVE TESTS
(Score 4 from Table 17)

[illegible]

$$= \frac{N^2 d_1 d_2 - (\Sigma f_{12})^2 (\Sigma f_{12})}{\sqrt{(\Sigma f_{12})^2 - (\Sigma f_{12})^2} \sqrt{N^2 d_1 d_2 - (\Sigma f_{12})^2}} = \frac{49(2,492) - (281)^2 (360)}{\sqrt{(2,492)^2 - (281)^2} \sqrt{13(3,300) - (360)^2}} = \frac{7716}{7716} = \frac{132(111)}{132(111)} = \frac{7716}{14,052} = 53$$

pretest scores across the bottom, from low to high, and the midterm scores in ascending order on the left side of the figure. Then, using the scores in Table 17, put a tally (/) in the proper cell for each pair of scores. Find the person's X score along the bottom of the scattergram and then go up in that column until you are directly opposite his Y score. Put a tally in that cell. Do this for each individual, and you will have a total of 43 tallies, for there are 43 pairs of scores.

For example, the lonesome-looking entry in the cell designating an X score of 30-41 and a Y score of 12-15 indicates the student who earned 40 points on the pretest but only 15 at midterm. The highest score on the pretest was 73, earned by the person who made 54 on the midterm exam, this is shown by the tally at the far right.

After the tallying is over, change the tallies to numbers.

Computing r from the scattergram. Now you are ready to find r . Set up the four new rows at the top of the scattergram and the six columns at the right. The first row is for f_x , the frequency of X scores in each column of the scattergram. These are 1, 1, $1 + 1 = 2$, $1 + 1 = 2$, $3 + 1 + 1 + 2 = 7$, and so forth. The first column to the right of the scattergram is for f_y , the frequency of Y scores in each row of the scattergram. These are, from top to bottom, $1 + 1 = 2$, $1 + 1 = 2$, $1 + 1 + 1 + 2 + 1 = 6$, and so on.

The second row at the top, set in bold-face type and labeled d_x , shows deviations from the arbitrary origin, which in this case is $\frac{30 + 32}{2} = 31$,

the midpoint of the lowest X class. These d_x values start at 0 and go by 1's up to 14. The d_y column at the right begins at the bottom with 0 and goes by 1's up to 13. Obviously, the numbers continue by 1's until the highest class is reached. How high they go depends upon the number of classes for the X and Y variables in the scattergram. The highest number for d_x will always be one less than the number of X classes, and the top number for d_y will always be one less than the number of Y classes.

The rest of the procedure is self-explanatory, if the meaning of the symbols is understood. The third (56-59) row of the scattergram will serve to show how the " $f_x \times d_x$ row sums" are obtained. There are frequencies in five different cells of that row. These are f_x 's for the row. To get the $f_x \times d_x$ row sum for this third row we take the 1 farthest left and multiply it by its d_x of 5. The next 1 is multiplied by its d_x of 7, the next 1 by 9, the 2 by 10, and the 1 farthest right by 12. $(1 \times 5) + (1 \times 7) + (1 \times 9) + (2 \times 10) + (1 \times 12) = 53$, the figure shown in the " $f_x \times d_x$ row sums" column.

There are two checks in the table. The sum of the f_x row at the top should equal the sum of the f_x column at the right, since both equal N , the number of pairs of scores. Also, the third row at the top sums to $\Sigma f_x d_x$, which is the same as the sum of the " $f_x \times d_x$ row sums" column at the right.

Because there are no other automatic checks, all computations should be gone over carefully, preferably by someone besides the original computer.

Substituting in the formula is straightforward, if one simply follows the symbols. Each of the square roots in the denominator should be carried out to four figures and rounded back to three for computational ease (see Appendix D, pages 456-458). You may find a table of square roots helpful at this point. The final r should not usually be reported to more than two figures, unless it is based upon several hundred individuals.

Computing the two means from scattergram data. On page 79 a formula for the mean was given as $M = M' + \frac{1}{N} \sum fd$. For the X scores in Table 18, the assumed mean, M' (or, more properly in this example, the *arbitrary origin*), falls at the middle of the 30-32 class, which is $(30 + 32)$ divided by 2, or 31. The formula for the mean of the X distribution is

$$M_x = M' + \frac{1}{N} \sum fd_x = \frac{30 + 32}{2} + \frac{3 \times 281}{43} = 31 + \frac{843}{43} = 50.6$$

Similarly, the mean of the Y distribution is

$$M_y = M' + \frac{1}{N} \sum fd_y = \frac{12 + 15}{2} + \frac{4 \times 360}{43} = 13.5 + \frac{1440}{43} = 47.0$$

Of course, this does not mean that the students knew less at the midterm than when they began the course. Had the Y test been given at the first of the quarter, the average score would probably have been only slightly above zero, since the midterm examination covered much more advanced material than the pretest.

Computing the two standard deviations from scattergram data. The formula for obtaining the standard deviation from grouped data that was explained on page 83 quickly yields standard deviations for the X and Y distributions, since the two square root values have already been obtained as the denominator, $(132)(111)$, in the Table 18 computation of r .

$$\sigma_x = \frac{1}{N} \sqrt{\sum f d_x^2 - (\sum f d_x)^2} = \frac{3 \times 132}{43} = \frac{396}{43} = 9.2$$

$$\sigma_y = \frac{1}{N} \sqrt{\sum f d_y^2 - (\sum f d_y)^2} = \frac{4 \times 111}{43} = \frac{444}{43} = 10.3$$

Determining the medians. It may be helpful at this point to review the computation of the median by obtaining this statistic for the X and Y distributions by the method explained on pages 76-77. The median is the 50th percentile. Fifty per cent (or half) of 43 is 21.5. I look at the first row at the top of the scattergram and count f_x frequencies from left to right: $1 + 1 + 2 + 2 + 7 + 3 = 16$. This leaves you "suspended" between the 4 classes 15-17 and 18-20 or at 17.5. Now count $21.5 - 16 = 5.5$ units into the 18-20 class which has a frequency of 7 and an interval of 3

(because it includes scores of 18, 19 and 50) $\frac{55}{7} \times 3 = \frac{165}{7} = 24$ and $175 + 24 = 199$ the median of the N (pretest) distribution. Check by counting the other way $505 - \left(\frac{15}{7} \times 3\right) = 199$. Compare this median with the mean which is 206.

The median of the N (midterm) examination is obtained similarly by way of the f_r frequencies in the first column at the right of the scattergram $135 + \left(\frac{65}{7} \times 1\right) = 172$. Check by counting down $475 - \frac{05}{7} \times 4 = 472$. This is about the same as the N mean of 470.

Rank correlation If the measures to be correlated are consecutive, untied ranks 1, 2, 3 and so on through N , the r formula reduces to $1 - \frac{6\sum D^2}{N(N^2 - 1)}$ where D represents the positive difference between paired ranks and N is the number of pairs. An illustration will show how simple the computational procedure is.

A certain test question requires that six historical events be ranked in chronological order, 1 representing the earliest and 6 the most recent without ties. Table 19 shows the events, the ranks assigned by Richard Roe and by John Doe, and the correct ranks. The first r is secured by summing the squared differences between Richard's ranks and the correct

TABLE 19

THE COMPUTATION OF RHO FOR EACH OF TWO STUDENTS WHO ARRANGED SIX HISTORICAL EVENTS IN CHRONOLOGICAL ORDER

Events to Be Ranked (1 = earliest)	Richard Roe's Ranks	Correct Ranks	D	D^2	John Doe's Ranks	Correct Ranks	D	D^2
French Revolution	2	6	4	16	5	6	1	1
Magna Charta	4	3	1	1	4	3	1	1
Pompeii Destroyed by Vesuvius	3	1	2	4	2	1	1	1
Columbus Discovers America	1	4	3	9	3	4	1	1
Fall of Roman Empire	5	2	3	9	1	2	1	1
Spanish Armada Destroyed	6	5	1	1	6	5	1	1
$N = 6 \quad \sum D^2 = 40$					$\sum D^2 = 6$			
$\rho \text{ (rho)} = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 40}{6 \times [(6 \times 6) - 1]}$					$\rho = 1 - \frac{6 \times 6}{6 \times 35}$			
$= 1 - \frac{40}{35} = 1 - 1.14 = -.14$					$= 1 - \frac{6}{35}$			
					$= 1 - .17$			
					$= .83$			

ranks and substituting in the formula. This r , usually called ρ and written as ρ , equals $-.14$. Like r , ρ can vary from -1 through 0 to $+1$. The negative ρ found for Richard probably indicates poor guessing rather than actual misinformation. Note that John Doe, with a ρ of $.83$, seems to have had a fairly good general knowledge of the correct chronology, even though he "missed" every one of the six ranks.

TABLE 20

THE VARIOUS VALUES OF ρ FOR ALL POSSIBLE SUMS OF SQUARED DEVIATIONS (ΣD^2) FOR N 'S FROM 2 THROUGH 10

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

Sum of Squared Deviations, ΣD^2	Number of Things Ranked (N)									Sum of Squared Deviations, ΣD^2
	2	3	4	5	6	7	8	9	10	
0	1 00	1 00	1 00	1 00	1 00	1 00	1 00	1 00	1 00	0
2	-1 00	50	80	90	94	96	98	98	99	2
4		0	60	80	89	93	95	97	98	4
6		-50	40	70	83	89	93	95	96	6
8		-1 00	20	60	77	86	90	93	95	8
10			0	50	71	82	88	92	94	10
12			-20	40	66	79	86	90	93	12
14			-40	30	60	75	83	88	92	14
16			-60	20	54	71	81	87	90	16
18			-80	10	49	68	79	85	89	18
20			-1 00	0	43	64	76	83	88	20
22				-10	37	61	74	82	87	22
24				-20	31	57	71	80	85	24
26				-30	26	54	69	78	84	26
28				-40	20	50	67	77	83	28
30				-50	14	46	64	75	82	30
32				-60	09	43	62	73	81	32
34				-70	03	39	60	72	79	34
36				-80	-03	36	57	70	78	36
38				-90	-09	32	55	68	77	38
40				-1 00	-14	29	52	67	76	40
42					-20	25	50	65	75	42
44					-26	21	48	63	73	44
46					-31	18	45	62	72	46
48					-37	14	43	60	71	48
50					-43	11	40	58	70	50
52					-49	07	38	57	68	52
54					-54	04	36	55	67	54
56					-60	0	33	53	66	56
58					-66	-04	31	52	65	58
60					-71	-07	29	50	64	60
62					-77	-11	26	48	62	62
64					-83	-14	24	47	61	64
66					-89	-18	21	45	60	66
68					-94	-21	19	43	59	68
70					-1 00	-25	17	42	58	70

TABLE 20 (Continued)

THE VARIOUS VALUES OF RHO (ρ) FOR ALL POSSIBLE SUMS OF SQUARED DEVIATIONS (ΣD^2) FOR N 'S FROM 2 THROUGH 10

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

ΣD^2	7	8	9	10	ΣD^2	7	8	9	10
72	-29	11	40	56	122	-45	-02		26
74	-32	12	38	55	124	-48	-03		25
76	-36	10	37	54	126	-50	-05		24
78	-39	07	35	53	128	-52	-07		22
80	-43	05	33	52	130	-55	-08		21
82	-46	02	32	50	132	-57	-10		20
84	-50	0	30	49	134	-60	-12		19
86	-54	-02	28	48	136	-62	-13		18
88	-57	-05	27	47	138	-64	-15		16
90	-61	-07	25	45	140	-67	-17		15
92	-64	-10	23	44	142	-69	-18		14
94	-68	-12	22	43	144	-71	-20		13
96	-71	-14	20	42	146	-74	-22		12
98	-75	-17	18	41	148	-76	-23		10
100	-79	-19	17	39	150	-79	-25		09
102	-82	-21	15	38	152	-81	-27		08
104	-86	-24	13	37	154	-83	-28		07
106	-89	-26	12	36	156	-86	-30		05
108	-93	-29	10	35	158	-88	-32		04
110	-96	-31	08	33	160	-90	-33		03
112	-100	-33	07	32	162	-93	-35		02
114		-36	05	31	164	-95	-37		01
116		-38	03	30	166	-98	-38	-01	
118		-40	02	28	168	-100	-40	-02	
120		-43	0	27					

Table 20 makes the computation of rho itself unnecessary when the number of pairs is less than 11. Just compute ΣD^2 , look in Table 20 for this figure, move over to the appropriate N column, and read the value of rho there. For example, the ΣD^2 for Richard Roe was 40. Looking in the ΣD^2 column at 40 and then to the right under the N of 6 we find that rho is $-.14$, which agrees with the value computed in Table 19.

A simplified scoring procedure for "sequence" items is described in Appendix C on page 455.

Strictly speaking, the rho formula is not appropriate when any ties occur. It is sometimes used as a short-cut procedure for estimating the r between scores by first changing the two sets of scores to ranks. If only an approximate measure of relationship is desired, this method may yield satisfactory results, despite ties. It will usually be tedious when N is as great as 30, however, for the ranking process will be time-consuming and the fractional ranks will be hard to square. A worked example appears in Table 21. It is based upon 20 pairs of scores in Table 7, page 67, for which the r from Table 8 is .65. Note that rho is .67, a discrepancy of only .02.

TABLE 20 (Continued)

THE VARIOUS VALUES OF RHO (ρ) FOR ALL POSSIBLE SUMS OF SQUARED DEVIATIONS (ΣD^2) FOR N 's FROM 2 THROUGH 10

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

ΣD^2	9	10	ΣD^2	9	10	ΣD^2	9	10
170	-42	-03	224	-87	-36	278		-68
172	-43	-04	226	-88	-37	280		-70
174	-45	-05	228	-90	-38	282		-71
176	-47	-07	230	-92	-39	284		-72
178	-48	-08	232	-93	-41	286		-73
180	-50	-09	234	-95	-42	288		-75
182	-52	-10	236	-97	-43	290		-76
184	-53	-12	238	-98	-44	292		-77
186	-55	-13	240	-100	-45	294		-78
188	-57	-14	242		-47	296		-79
190	-58	-15	244		-48	298		-81
192	-60	-16	246		-49	300		-82
194	-62	-18	248		-50	302		-83
196	-63	-19	250		-52	304		-84
198	-65	-20	252		-53	306		-85
200	-67	-21	254		-54	308		-87
202	-68	-22	256		-55	310		-88
204	-70	-24	258		-56	312		-89
206	-72	-25	260		-58	314		-90
208	-73	-26	262		-59	316		-92
210	-75	-27	264		-60	318		-93
212	-77	-28	266		-61	320		-94
214	-78	-30	268		-62	322		-95
216	-80	-31	270		-64	324		-96
218	-82	-32	272		-65	326		-98
220	-83	-33	274		-66	328		-99
222	-85	-35	276		-67	330		-100

Interpreting the coefficient of correlation. In interpreting the coefficient of correlation, two things must be considered. The first is the *sign* of the coefficient. The sign indicates the *direction* of the relationship. Positive coefficients indicate direct relationship, that is, there is a tendency for the two series of values to vary in the same direction, high values in one column being associated with high values in the other column, low values in one column being associated with low values in the other column, and so on. On the other hand, negative coefficients indicate inverse relationship, that is, there is a tendency for the two series of values to vary in opposite directions, high values in one column being associated with low values in the other column, and high values in that column being associated with low values in the first column.

Another thing is equally important and far more difficult to interpret, that is the *magnitude* or *size* of the coefficient. The *size* of the coefficient indicates the *degree* or *closeness* of the relationship, just as the *sign* of the

TABLE 21

ESTIMATING THE COEFFICIENT OF CORRELATION BY THE
SIFARMAN RANK DIFFERENCE METHOD

Scores		Ranks		Differences in Ranks	
Educational Age (E A)	Mental Age (MA)	EA	MA	D	D ²
188	208	1	2	1	1
186	218	2	1	1	1
185	201	3	3	0	0
183	185	4.5	8.5	4	16
183	165	4.5	17	12.5	156.25
182	191	6	6	0	0
181	185	7	8.5	1.5	2.25
180	193	8	5	3	9
179	181	9	10	1	1
176	165	11.5	17	5.5	30.25
176	187	11.5	7	4.5	20.25
176	176	11.5	14	2.5	6.25
176	180	11.5	11.5	0	0
175	166	14	15	1	1
174	197	15	4	11	121
173	154	16	20	4	16
171	180	17	11.5	5.5	30.25
167	163	18.5	17	1.5	2.25
167	177	18.5	13	5.5	30.25
165	164	20	19	1	1

 $\Sigma D^2 = 445$ $N = 20$

$$r = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 445}{20[(20)^2 - 1]} = 1 - \frac{2670}{7980} = 1 - .33 = .67$$

coefficient indicates the *direction* of the relationship. The minimum coefficient is .00, which indicates no consistent relationship whatsoever. From this minimum value the coefficients increase in both directions until -1.00 is reached for one limit, and 1.00 for the other. It should be noted that both -1.00 and 1.00 indicate equally close relationship, for both are perfect. Their one important difference is in direction, the former being inverse and the latter being direct. In like manner, all other values of the same size, such as $-.50$ and $.50$, indicate equally close relationship. It is the size, not the sign of the coefficient that gives the clue to the closeness or degree of relationship.

The problem, then, is to know how close a relationship is indicated by a coefficient of correlation of a given magnitude, regardless of sign. For example, how close a relationship is indicated by a coefficient of .60? Unfortunately, there is no simple way of answering such a question. Attempts to indicate this relationship by some descriptive adjective, such as "high" or "marked," are vague and often misleading, to say the least. As a matter

of fact, a coefficient of .60 might be regarded as high for one type of situation and low for another. For example, a coefficient of .60 between a general intelligence test administered at the beginning of the year and school marks recorded at the end of the year might be regarded as high, because such coefficients usually fall below that. But a coefficient of .60 between scores on two forms of this intelligence test administered the same day would be unusually low. In other words, "high" and "low" have only *relative* meaning. Before an interpretation can be made of a coefficient on this basis, the reader must at least know what the central tendency of such coefficients for similar data is.

Expectancy tables. A most helpful way to interpret the degree of relationship between two variables is to construct an *expectancy table* from the scatter diagram and inspect it carefully. This may be done with the Table 18 data on page 92 by calling scores above 50 on the pretest and scores higher than 47 on the midterm exam "above average," as shown in Table 22. These cutting points give as near a 50 per cent split of the scores

TABLE 22

A SIMPLE EXPECTANCY TABLE BASED UPON THE 43 PAIRS OF SCORES IN THE TABLE 18 SCATTERGRAM FOR WHICH $r = .53$

	<i>Below Average on Pretest (30-50)</i>	<i>Above Average on Pretest (51-74)</i>	<i>Sums</i>
<i>Above Average at Midterm (48-67)</i>	$\frac{7}{23} = 30\%$	$\frac{14}{20} = 70\%$	21
<i>Below Average at Midterm (12-47)</i>	$\frac{16}{23} = 70\%$	$\frac{6}{20} = 30\%$	22
<i>Sums</i>	23	20	43

as possible. Twenty three persons are "below average" on the pretest and 20 "above average." Twenty two students are "below average" on the midterm examination and 21 "above average."

Of those students who were below average on the pretest, 70 per cent were also below average on the midterm examination, so the odds are 7:3 that a person scoring below average initially will six weeks later again score below average. The same 7:3 odds hold for those who score above average on the pretest: this will not necessarily be precisely true in other similar problems because the number of persons "above average" may differ considerably for the two tests.

The 7 individuals who went from below average on the pretest to above average at midterm might be called "false negatives," since at first they were classified too low while the 6 who went from above average to below

average might be labeled 'false positives' Only 3 of the 7 false negatives changed greatly, the other 4 were not very low on the pretest or very high at midterm Likewise, none of the 6 false positives were very high on the pretest or very low at midterm Obviously though the four cell expectancy table provides a useful summary it is less exact than the scattergram¹⁵

Validity and reliability coefficients One of the most important uses of the coefficient of correlation is in determining the validity of a test There are two types of validity, or rather two methods of judging the validity of a test namely, curricular and statistical The former is subjective and the latter is objective *Curricular validity* is determined by examining the content of the test itself and judging the degree to which it is a true measure of the important objectives of the course, or a truly representative sampling of the essential materials of instruction *Statistical validity* is determined by setting up a criterion of the thing which it is desired to measure and then computing the coefficient of correlation between the test scores and the criterion The r so obtained is called a *validity coefficient*, and is interpreted like any other coefficient of correlation

A second use of the coefficient of correlation is in determining the reliability of a test Since reliability is the degree of consistency with which the test measures whatever it does measure several ways to determine reliability are to be found in computing the coefficient between two forms of the same test two matched halves of a test or two administrations of the same test

The r discussed in this section called the zero-order coefficient of correlation, is only one of many different types of correlation coefficients

G Measures of Error

Errors in educational measurement may be grouped conveniently into three types according to source

- 1 Errors of technique
 - a Arithmetical errors in computation or the like
 - b Use of inappropriate measures
- 2 Errors of measurement
 - a Imperfect measuring instruments
 - b Unskilled tester
 - c Fluctuations in the persons measured
- 3 Errors of sampling
 - a Selection or bias in sampling
 - b Chance fluctuations in random sampling

¹⁵ A very clear discussion of various types of expectancy tables of which the one above is the simplest is contained in Alexander G Weeman Expectancy Tables—A Way of Interpreting Test Validity Test Service Bulletin No 38 15 December 1949 copies of which may be obtained free from the Psychological Corporation 32 Fifth Avenue New York 18 New York

Errors of technique. Obvious types of errors are mistakes in adding scores and various computational errors in statistical analysis. The only protection against such errors is the exercise of great care. Likely to be more serious are the errors due to the use of inappropriate measures for the data in hand. It is poor technique to introduce more refined measures than the data warrant or the purpose requires. All statistical formulas are based upon certain assumptions which often are not fully met in actual practice. The following are common examples. Computations based upon data in a grouped frequency distribution depend for complete accuracy on the assumption that the scores are uniformly distributed within the several intervals or that the midpoint of each interval may be used to represent the average value of all scores in the interval. The Pearson r assumes linear correlation between the two variables—constancy of the relationship throughout the range of scores. Many formulas are based on the assumption of a normal distribution of the measures. Whenever the data in a given situation fail to conform to these assumptions, certain errors are introduced. Fortunately, in actual practice, these errors are often not great enough to introduce serious errors of interpretation. But gross errors due to the use of inappropriate techniques are sufficiently numerous to warrant extreme caution. Turfey and Daly¹⁶ made a study of articles containing product-moment r 's and came to the conclusion that this technique is employed "with little regard to the fulfillment of the necessary antecedent conditions." In fact, in 60 of the 63 articles studied they found that "their authors have left themselves open to the suspicion of having employed the correlation technique in a way which is meaningless, if not positively misleading."

Errors of measurement. There are many possible sources of errors in measurement, even when there are no computational errors and when the most appropriate statistical analysis has been employed. In the first place, no measuring instrument is perfectly valid or perfectly reliable. In the second place, the personal equation of the examiner must be reckoned with. Inexperienced examiners may allow too much or too little time in administering the test, or may otherwise depart from standardized procedure in administering the test or in scoring the papers. In the third place, there is likely to be great variability in the responses of the subjects taking the test. Accidental occurrences, such as the breaking of a pencil point on timed tests, fluctuations in motivation, fatigue, and other physical and mental factors may seriously affect the test results.

It will be noted that some errors of measurement are systematic and tend to affect all individuals alike. Allowing too much or too little time on a test of reading speed is an example. On the other hand, many errors of a variable character occur affecting the individuals unequally or in dif-

¹⁶ Paul H. Turfey and Joseph F. Daly. Product-moment Correlation as a Research Technique in Education. *Journal of Educational Psychology* 26: 200-211, March 1935.

ferent directions. Sensory defects, health conditions, and motivation are examples of conditions that produce variable errors in measurement. The effects of these errors are presented briefly in Table 23.

TABLE 23
EFFECTS OF CONSTANT AND VARIABLE ERRORS ON CERTAIN TYPES OF STATISTICS

Measure	Constant Errors	Variable Errors
Central Tendency	Increased or decreased by amount of the error	Usually tend to offset or balance each other
Variability	Little or no effect	Usually made too large
Relationship	Little or no effect	Usually made too small

It will be observed that constant errors affect measures of central tendency most seriously, and often there are no methods for correcting this bias.

Errors of sampling. It is usually impractical to measure all the cases of a given type. For example, it would be a formidable task to obtain the mean IQ of all high-school freshmen in a state, or the difficulty of each word in a series of textbooks. Fortunately, it is not necessary to do so. It has been found possible to estimate the range within which the true measure probably lies. But to do so, one needs a representative sampling of the total population. Against errors in a selected or "hand-picked" sampling there is no statistical protection. An adequate sampling may be chosen in a random manner, and the larger the sampling, the better, although increasing the number of cases does not in itself eliminate the possibility of error. The sampling method determines whether or not a biased (non-representative) sample will be obtained.

Further reading. In a stimulating article entitled "Errors, Estimates, and Samples—the Indispensable Concepts," Charles R. Langmuir¹⁷ makes clear many points that have only been hinted at in this chapter. Three other sources of valuable supplementary information are "Making Test Scores Meaningful,"¹⁸ "The Three-Legged Coefficient,"¹⁹ and "Reliability and Confidence."²⁰ They will do much to help the statistically untrained teacher or administrator understand important aspects of measurement theory that might otherwise remain vague.

¹⁷ Arthur E. Traxler (Editor), "Measurement and Evaluation in the Improvement of Education," *American Council on Education Studies*, 15: 68-81, Series I, No. 48, April 1951, 1785 Massachusetts Ave., N.W., Washington 6, D.C.

¹⁸ William B. Schradler, *College Board Review*, No. 14: 202-208, May, 1951, 425 West 117th St., New York 27, New York.

¹⁹ Alexander G. Wesman, *Test Service Bulletin* No. 40: 1-3, December 1950, Psychological Corporation, 522 Fifth Ave., New York 36, New York. Free.

²⁰ Alexander G. Wesman, *Test Service Bulletin* No. 44: 1-6, May 1952, Psychological Corporation. Free.

H. Summary

The following is an outline summary of three important concepts and some statistics useful in connection with test scores and other quantitative data

1 Central tendency

a The *mean* often called the 'average' in everyday life, is obtained by summing all the scores and dividing this sum by the number of scores. It is for most purposes the best measure of central tendency.

b The *median* is a point above which half of the scores lie and below which the other half lie. Thus it is the 50th percentile or Q_2 .

c The *mode* is the most frequent score a rather crude measure.

2 Variability

a The *standard deviation* (SD or σ) involves every measure in the distribution. Approximately two thirds of all scores in a "normal" frequency distribution lie not more than one standard deviation away from the mean.

b D , a percentile measure of dispersion is the distance between the 90th and 10th percentiles. Four tenths of D ($0.4D$) provides a fairly good estimate of the standard deviation.

c Q , the quartile deviation or semi interquartile range, is half of the distance between Q_3 (the 75th percentile) and Q_1 (the 25th percentile). Though widely used Q is usually a poorer measure of variability than D , which in turn is somewhat inferior to σ .

d The *range* is the distance from the lower real class limit of the lowest class to the higher real class limit of the highest class. For most purposes it is a very inadequate measure of variability.

3 *Correlation* *covariation*, or *concomitant variation*. There are numerous ways of expressing co-relationship. The most common statistic is Pearson's r . A simplification of r , applicable chiefly to data originally secured in the form of ranks is rho (ρ). Both r and ρ have values of -1 for perfect inverse relationship, 0 for sheer chance association, and $+1$ for perfect direct relationship.

Reliability and validity coefficients are usually r 's secured under certain special conditions.

I Instructional Test Items

Appendix A, pages 429-435 contains 50 five-option multiple-choice items covering the material in this chapter. After you have gone through Chapter 3 carefully turn to them and test your knowledge. Refer to the chapter as much as you please.

SELECTED REFERENCES FOR FURTHER READING

Chamberlain, L. G. *Statistical Calculation for Beginners* (Second Edition). Cambridge: Cambridge University Press, 1932. 168 pages.

- Dixon, Wilfred J, and Massey, Frank J Jr *Introduction to Statistical Analysis* New York McGraw-Hill Book Company, 1951 370 pages
- Edwards, Allen L, *Statistical Analysis for Students in Psychology and Education* New York Rinehart & Company, 1946 360 pages
- Garrett, Henry F, *Statistics in Psychology and Education* (Fourth Edition) New York Longmans Green & Company, 1953 460 pages
- Guilford, J P *Fundamental Statistics in Psychology and Education* (Second Edition) New York McGraw-Hill Book Company 1950 633 pages
- Indquist E F *A First Course in Statistics* Boston Houghton Mifflin Company, 1942 242 pages
- Odell, C W, *An Introduction to Educational Statistics* New York Prentice-Hall Inc, 1946 269 pages
- Tippett, L H C, *The Methods of Statistics* (Fourth Edition) New York John Wiley & Sons 1952 395 pages
- Walker, Helen M, *Elementary Statistical Methods* New York Henry Holt & Co, 1943 368 pages
- Walker, Helen M, and Lev, Joseph *Statistical Inference* New York Henry Holt and Company, 1953 510 pages
- Yule, G Udny, and Kendall Maurice G *An Introduction to the Theory of Statistics* New York Hafner Publishing Company, 1950 701 pages

4

The Characteristics of a Satisfactory Measuring Instrument

A. Introduction

Importance of the problem. What are the earmarks of a good test, examination, or other measuring instrument? In the selection of a test, as in the selection of an automobile, it is important to know what to look for. There is usually a choice among many possibilities which are very unequal in merit. Each year many automobiles are bought because of the appeal of some gadget, such as a fancy radiator ornament or cigarette lighter, and many standard tests are bought for no better reason. Whether a purchaser *buys*, or is merely *sold*, depends largely on whether or not he knows what to look for in the article in question. Moreover, every teacher will have occasion to use tests of his own construction, and should know what qualities to strive for in such tests. As a rule, the same characteristics are essential in an informal test made by the classroom teacher as in a standard test bought ready-made from a publisher.

In any satisfactory measuring instrument three qualities are indispensable. These are

- 1 Validity
- 2 Reliability
- 3 Usability

It is essential, therefore, that every teacher have a clear idea regarding the meaning of these characteristics, and know how to judge their presence in tests whether standardized or nonstandardized.

B. Validity

Meaning of validity. One kind of validity concerns the degree to which the test or other measuring instrument measures what it claims to. In a word, *validity* means *truthfulness*.¹ Does the test really measure what it purports to? For example, whether a so-called "arithmetic reasoning test" is valid or not depends upon the extent to which it succeeds in measuring reasoning ability in arithmetic rather than other things, such as reading ability or general intelligence. Validity, then, refers to the truthfulness of the test and is always its most important characteristic. No matter what other merits the test may possess, if it lacks validity, it is worthless. Whether you are selecting a standard test or making an informal test, the first thing to consider is its validity. How, then, does one judge whether or not a test or other measuring instrument is valid?

General considerations. The answer to this question may best be approached by giving attention to some preliminary considerations of a general nature.

1 The *nature of the thing being measured* must always determine the methods and materials of measurement. In order to judge the validity of an intelligence test, for example, it is necessary to consider what intelligence is, what its qualities are, or at any rate, how it manifests itself. In like manner, in order to judge the validity of an achievement test, it is necessary to consider what it is that the achievement test is supposed to measure. This means that the first step in judging the validity of achievement tests is a clear statement of the specific objectives of the course or subject.

2 Any measurement in education is always a *sampling* never entirely complete. The test maker relies upon a sample much as does the chemist in the health department in passing upon the quality of the city's water supply. In psychological language, any test is merely a series of situations designed to call forth a sufficient number of representative responses to enable the examiner to determine the amount of the thing in question that happens to be present.

3 The *accuracy of the measurement*, its fineness of discrimination, will depend upon the purpose it is to serve. A cheap alarm clock will usually suffice for a housewife in determining when to prepare lunch or to expect the postman, but a finer timepiece is required for the locomotive engineer. In like manner, a sundial or hour glass may be adequate for a gardener but a split-second watch is essential for a football official. It would be almost as absurd to attempt to use a sundial to time a football game as to use it for measuring temperature or wind velocity. In other words, the validity of the measuring instrument must always be considered in relation to the purpose it is to serve. Validity is always specific, in relation to some definite situation. A test is not just valid, it is *valid for something*. There is no such thing as general validity.

¹ See page 111 for the distinction between curricular and statistical validity.

I The Validation of Intelligence Tests

Although the job of constructing the so-called tests of general intelligence is usually turned over to the specialist, a general knowledge of how such tests are validated will enable the teacher to select and use them more discriminatingly

The meaning of intelligence. What, then, is meant by "general intelligence," the thing such tests claim to measure? Although there is no unanimity among psychologists regarding the exact definition of intelligence, there is substantial agreement that what existing tests attempt to measure is capacity to learn, particularly to learn the academic tasks imposed by the school. Such a conception of intelligence is not very "general" after all. It is clearly narrower than the popular notion, since it is restricted largely to abstract intelligence and leaves out of account social intelligence, mechanical intelligence, and intelligence in special fields such as athletics, music, or oratory.

It is also clear at the outset that intelligence can be measured only indirectly, its presence must be inferred from the observed behavior of the individual, his reactions to certain carefully chosen and controlled situations called tests. Such tests should meet two general requirements. First, there must be a sufficiently large and varied assortment of test situations to call forth a wide variety of mental operations, primarily of the higher type, such as imagination, judgment, and reasoning. Second, the situations must be of such a nature that every individual taking the test has had approximately equal opportunity to learn, and as far as possible, equal motivation. This second standard is hard to meet and is usually only approximated even in the best tests. It clearly rules out tests that involve special talents such as for music or art, and makes questionable those that depend on specific school experience, which is by no means uniform for all pupils. In general, group tests meet these standards especially the second, less well than do individual tests. The Army Alpha, for example, not only employs such situations as reading vocabulary and arithmetic, but, being originally designed for soldiers, has material that is more within the experience of men and boys than of women and girls.

The Terman criteria. In developing the 1916 Stanford Revision of the Binet Scale, Terman relied upon three additional criteria of intelligence namely, age increase, coherency, and world success.² *Age increase* means that each test item must show an increasing percentage of successful responses from one year level to the next. This is only a partial criterion, since it must assume that the items chosen are of a type that may reasonably be expected to measure intelligence. Purely physical measurements, for example, such as strength of grip, or speed in running, show age increases. The second

² For a discussion of the procedure used in the Revised (1937) Stanford Binet see Lewis M. Terman, 'The Revision Procedures' in Quinn McNemar, *The Revision of the Stanford Binet Scale*, pp. 1-14. Boston: Houghton Mifflin Company, 1942.

criterion *coherency*, is based on the assumption that the whole test is a more valid measure of intelligence than are any of its parts. Upon the basis of the entire test the group is divided into dull, normal, and bright sections. Then to be acceptable each item must discriminate among the sections by showing a progressively increasing percentage of successes as we go from dull to bright. This procedure really measures the internal consistency of the test, much as a logician judges the validity of a course of reasoning. Both Galton and Binet used the method of contrasting groups although their groups were selected from external criteria rather than from the test itself.

The third criterion, *world success*, is the ordinary common sense standard of everyday life. As the test is validated on children this really means the child's world, which is primarily that of the school, his standing in which is reflected in his academic record. This is of course not a perfect criterion. It is not only highly subjective in character but it throws the primary responsibility ultimately upon the judgment of teachers which because of its limitations the test is being designed to replace or supplement. This is not so bad as it seems however for the basis is not that of the pupil's mark on a single examination whose notorious unreliability has already been described but rather that of his *entire record* for an extensive period, a far more stable thing. Furthermore the reliance is usually placed not upon the judgment of any single teacher but rather upon the average of several experienced teachers. The consensus of competent persons is the ultimate criterion of values from the constitutionality of a law down to the beauty contest at the local theater.

Individual versus group tests. It is generally assumed that an individual test is likely to meet more fully the criteria described than does a group test. Furthermore the individual test permits the trained examiner to observe more carefully the behavior of the subject during the course of the examination. For example if the subject shows signs of nervousness or refuses to co-operate fully the examiner realizes that a valid measure of intelligence is impossible under the circumstances and so waits for a more opportune time. Also if the subject is handicapped by defective vision or hearing this condition is likely to be discovered by the examiner who then takes it into account in making his interpretations and recommendations. For these reasons the individual intelligence test is usually taken as the criterion or standard for validating the group test. However sometimes a group test is validated by comparing the scores made on that test with those made by the same individuals on another group test or possibly some combination of two or more group tests. For all such comparisons with a criterion whether it be the Revised Stanford Binet or some group test or tests the Pearson product moment coefficient of correlation is usually employed. r is then referred to as the *validity coefficient*. If the agreement with the criterion is perfect the coefficient is 1.00 and if there is no

consistent relationship at all, the coefficient is .00. Naturally the nearer the coefficient approaches 1.00, the higher the validity is said to be, although in the last analysis everything depends upon the appropriateness of the criterion itself. Usually the most difficult step in test validation is securing an adequate criterion.

TABLE 24

INTERCORRELATIONS OF INTELLIGENCE TEST SCORES AND FIVE-SEMESTER AVERAGE GRADES FOR 284 SENIORS (124 BOYS, 160 GIRLS) IN A LOS ANGELES HIGH SCHOOL¹

Intelligence Test	Intelligence Test										Average Grade	
	Otis Self Administering		Terman McNemar		California Short-Form		SRA Non-Verbal		SRA Primary Mental Abilities			
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Otis	—	—	75	77	73	67	36	24	57	57	36	44
Terman McNemar	75	77	—	—	70	70	27	36	56	45	45	52
California Short-Form	73	67	70	70	—	—	31	24	54	56	38	50
SRA Non-Verbal	36	24	27	36	31	24	—	—	33	40	23	40
SRA PMA	57	57	56	45	54	56	33	40	—	—	42	44
Mean IQ	103.7		100.5		114.2		118.2		96.4		—	

Table 24 contains the intercorrelations of five intelligence tests and their correlations with average grades, separately by sex. It also shows the mean IQ on each test for the boys and girls lumped together. The Otis Self-Administering Higher Examination, Form A, correlates with the Terman-McNemar .75 for boys and .77 for girls, while the Terman-McNemar correlates only .27 and .36 with the Science Research Associates (SRA) Non-Verbal. Apparently the Otis, Terman-McNemar, and California Short-Form Test of Mental Maturity (1912 edition) are more closely related to each other than to the other two tests.

The most valid test for "predicting" the five-semester average-grade criterion is the Terman-McNemar, with coefficients of .45 for boys and

¹ These figures were secured from the following unpublished (mimeographed) report: Walter G. Heil and Alice Horn, *A Comparative Study of the Data for Five Different Intelligence Tests Administered to 284 Twelfth Grade Students at South Gate High School—Los Angeles*. Los Angeles: Curriculum Division, Los Angeles City School Districts, February 1950. 23 pages.

52 for girls, the SRA Non-Verbal is least valid. Note that in each of the five instances the r for girls is higher than for boys: 36 and 44, 45 and 52, 38 and 50, 23 and 40, and 42 and 44. These are not genuine validity coefficients, however. All 284 students were tested at the beginning of the last semester of their senior year, and the average grades are for the five previous semesters of high school work. Thus there is no prediction here from test scores to later school achievement as is implied by the term "validity coefficient."

Three such r 's are reported by Heil and Horn: though 19 between intelligence-test scores of 78 persons tested in the first semester of the first grade and their average high school marks, 36 for 54 individuals tested in the third grade, and 31 for 72 sixth graders. These validity coefficients are not high, but neither are some of the figures in the average grade column of Table 21, which are based upon an average interval of only 2½ semesters instead of 6-11½ years.

It is important to realize that scores on two tests may correlate perfectly even though the means are quite different. In other words, the size of r is wholly independent of the difference between means, so it tells us nothing about how interchangeable scores on two highly correlated tests are. In Table 21, the mean IQs for the same 284 students go from 96.4 for the Primary Mental Abilities total score to 118.2 on the SRA Non-Verbal, a difference of 21.8 IQ points! Yet both of these tests are issued by the same publisher. On the other hand, the Otis and Terman McNemar tests differ by only 1.8 points.

II The Validation of Achievement Tests

Curricular versus statistical validity. In some respects the validation of an achievement test is more difficult than the validation of an intelligence test, and a greater number of procedures are employed for its determination. In discussing the validation of achievement tests a distinction should be made between *curricular* validity and *statistical* validity. By curricular validity is meant the extent to which the content of the test is truly representative of the content of the course. Curricular validity implies an act of judgment as to the adequacy of the sampling included in the test. In the earlier days this was interpreted to mean merely the extent to which the items of the test included a representative sampling of the essential materials employed in instruction. More recently, however, curricular validity is thought of not primarily in terms of subject matter which at best is merely the stimulus, but rather in terms of the *mental reactions* expected of the pupils themselves.⁴ In other words, the center of gravity has shifted from the curriculum to the child.

⁴ Phillip J. Rulon, "On the Validity of Educational Tests," *Harvard Educational Review* 16: 290-296, October 1946. Reprinted by World Book Company as Test Service Notebook No. 3.

Statistical validity refers to the mathematical processes for determining the degree to which the test agrees with, or correlates with, some criterion which is set up as an acceptable measure of the thing in question. Some of these statistical procedures aim at validating the test as a whole and others at validating the items individually. Although the procedures commonly employed by professional test makers are often rather technical, especially for item validation, the essential ideas are relatively simple.

The technique employed in the preparation of the Cooperative Achievement Tests represents an effective combination of statistical analysis and the judgment of experts. These tests are constructed by a trained staff working in close co-operation with classroom teachers, subject-matter specialists, and test technicians. The procedure is outlined as follows:⁵

- a Preliminary planning and selection of content
 - Analyses of curricula, textbooks, research studies, etc
 - Formulation of objectives and determination of general plan
 - Preparation of detailed test outlines based upon survey of materials
 - Submission of outlines to authorities for criticism
 - Revision of test outlines in accordance with suggestion of critics
- b Preparation and editing of test items
 - Writing of items by test editors and cooperating experts
 - Submission of items to authorities for criticism
 - Revision of items in view of suggestions received
 - Preparation of experimental forms of test
- c Administration of experimental forms to a representative sampling of students to obtain item difficulty and validity indices, and to detect items which may be weak or ambiguous
- d Preparation of final form
 - Selection and revision of items for tentative final form
 - Obtaining from experts in subject matter fields, test technicians, etc., suggestions and criticisms of the tentative final form
 - Revision and final editing of the test, based on the criticisms and suggestions received
- e Administration of final form of test with earlier forms for equating and determination of scaled scores

Perhaps attention should also be called to certain limitations of frequency of mention or use as a criterion for selecting materials either for the curriculum or for the test. In the first place, to accept *what is* as a criterion of *what ought to be* leaves no room for progress. For example, someone has defined a synonym as a word you use when you do not know how to spell the word you want. It can scarcely be doubted that there is a wide margin between the words actually used in ordinary speaking and writing and those that should be used to convey best the meaning intended. In the second place, frequency by its very nature is a poor standard for judging importance. For example, birth and death occur but once in the life history

⁵ *Cooperative Achievement Tests for High School and College Classes*, page 5. New York: Cooperative Test Service, 1915. These tests are now distributed by the Cooperative Test Division of Educational Testing Service, 20 Nassau Street, Princeton, New Jersey.

of an individual, and yet who would say they are for this reason less important than dressing and undressing which occur every day? Frequency of use, therefore, although doubtless important as one measure of social utility, can rarely be regarded as the best criterion for validating a test. It should usually be employed with other criteria, rather than alone.

Some criticisms of test validity. One of the commonest criticisms of the validity of achievement tests, especially those of the objective type whether standardized or nonstandardized is that they are predominantly factual in character. It is alleged that they succeed merely in measuring verbal memory as distinguished from genuine understanding and leave unmeasured the really important outcomes such as discrimination judgment intellectual and emotional attitudes appreciations and the ability to make intelligent application of knowledge to new situations. Even the best friends of achievement tests will readily admit that as such tests are commonly made and used, the criticism has some merit. In fact no one has recognized the limitation of existing tests more clearly than some of the outstanding leaders of the measurement movement itself. Long ago Thorndike wrote ⁶

In the elementary schools we now have many inadequate and even fantastic procedures parading behind the banner of educational science. Alleged measurements are reported and used which measure the fact in question about as well as the noise of the thunder measures the voltage of the lightning. To nobody are such more detestable than to the scientific worker with educational measurements.

Thirteen years later Monroe⁷ wrote about the "child like faith in the efficacy of objective tests as instruments for measuring school achievement on the high school and college levels. Three examples of unwarranted beliefs were cited

I Objectivity in scoring is an essential requirement for a satisfactory test and if a test is objective the scores yielded by it may be considered highly accurate measures of school achievement.

II If a test has been shown to be highly reliable the scores yielded by it are highly accurate measures of the achievement specified by its announced or implied function.

III A high correlation with a criterion is sufficient evidence to justify the use of the scores yielded by a test as highly accurate measures of the achievement considered to be defined by the criterion.

While all three points are related to validity the first two are more appropriately treated in later sections. The third point merits further discussion here.

Validity coefficients are no exception to the general principle which holds that all coefficients of correlation are definitely influenced by the variability

⁶ E. L. Thorndike, *Measurement in Education*, *Twenty First Yearbook of the National Society for the Study of Education, Part I*, page 8. Quoted by permission of the Society, Bloomington, Illinois: Public School Publishing Company, 1922.

⁷ Walter S. Monroe, *Hazards in the Measurement of Achievement*, *School and Society* 41: 48-52, January 12, 1935.

the possession of knowledge and the ability to use it, which averages even less, one would hardly expect to find the perfect correlation between educational facilities and educational performance that such practice appears to assume. It is doubtless still true that there are Mark Hopkinses capable of transforming mere logs into colleges while marble palaces may remain but piles of stone for lack of such a magic touch. In any case, it is *safer* to examine what is happening to the student at his end of the log, than to remain content to measure the dimensions of the log, or even the credentials of the individual who happens to be at the other end.

Evidence concerning the value of indirect measures is conflicting. For example, in a study involving a test administered to 300,000 young men which measured both intelligence and general achievement, Davenport and Remmers¹⁰ found rather high r 's between the test means for states and such state characteristics as telephones per thousand persons (.83), per capita income (.81), value of school property (.76), and Negroes per thousand persons ($-.70$). They located and named "state economic," "rural-urban," and "deep-South versus non-South" factors which seemed to account for most of the correlations found. Their conclusion is¹¹

These data are all state data, they do not apply to individuals. Without much facetiousness, however, we interpret these results to mean that the probabilities of reaching a high educational achievement are much greater if one comes from a high income state which is highly urban, which is not in the South, and which has such advantages as library service available to most of its population, has a high proportion of foreign born citizens, a large number of residents in *Who's Who*, and many telephones.

Using achievement and intelligence test results from 154 communities, large and small, throughout the United States, Thorndike¹² found 24 community variables to be much more highly correlated with intelligence than with achievement. In fact, the estimated maximum correlation of these community aspects (population in thousands, per cent native-born white, and so forth) with achievement test means was only about .30. Thorndike offers several possible explanations of this low relationship.

It may, of course, be expedient at times to rely upon indirect evidence, but at any rate, one should do so only where direct evidence is not available and even then with full realization of the risks involved. For example, up to the present time test makers have found it difficult to devise suitable instruments for measuring such intangible outcomes of teaching as atti-

Kelley, *Interpretation of Educational Measurements*, page 19c. Yonkers-on Hudson: New York: World Book Company, 1927.

¹⁰ K. S. Davenport and Hermann H. Remmers, 'Factors in State Characteristics Related to Average A 12 V 12 Test Scores', *Journal of Educational Psychology* 41: 110-115, February, 1950.

¹¹ *Ibid.*, page 115.

¹² Robert L. Thorndike, 'Community Variables As Predictors of Intelligence and Academic Achievement', *Journal of Educational Psychology* 42: 321-338, October, 1951. Note of Correction: 43: 179-180, March 1952.

tudes appreciations and interests. But it is probably true that the better standard tests come far closer to measuring the objectives actually attained or aimed at in educational practice than they come to measuring those suggested as desirable in educational theory. It is apparently just as difficult to teach these intangible things as it is to test them. It seems reasonable to think that it is no less difficult to provide the appropriate teaching materials for bringing about the right kind of attitudes, appreciations and interests than it is to provide the appropriate testing materials for determining how well the job is being done. It should be kept in mind that a valid test consists largely of a representative sampling of the materials that make up the course. It should help to clarify the atmosphere once and for all to recognize frankly that the less tangible outcomes are harder to teach and to test than the more tangible outcomes. And it may be that for some time to come we shall have to be content to aim at both indirectly.

But this is no permanent solution to the problem. One of the important services the measurement movement can render education is the clarification of its objectives. The necessity for this should be apparent both to the curriculum maker and to the test maker. Considerable progress has already been made in this direction and more will doubtless be forthcoming. The pioneering work of Wrightstone¹³ is an illustration. He reports a series of tests in the social studies with such a diversity of aims as the interpretation of facts, the making of generalizations, the organization of data, several important work study skills and certain civic attitudes and beliefs. Reports of the Eight-Year Study of the Progressive Education Association indicate substantial progress in this direction.¹⁴

Item analysis. Specialists in test construction not only attempt to validate the test as a whole against some outside criterion but also to validate the items on the test individually, usually against an inside criterion, the test as a whole. Frequently an outside criterion would be better but it is often not available. Although many of the processes for item analysis are rather technical and complicated, the essential idea is easy to grasp. The purpose is to determine the difficulty and the discriminating value of each item in the test. Obviously an item missed by everybody or answered correctly by everybody who took the test is of no value in differentiating between good and poor pupils. If the test is for the purpose of determining the extent to which the minimum essentials of a unit or of a course have been mastered, however, the difficulty of the individual items is relatively unimportant and the matter of discrimination is of minor significance. But if the test is to be used over several grades as a basis of classification or

¹³ J. Wayne Wrightstone, "Measuring Some Major Objectives of the Social Studies," *School Review* 43: 771-779, December 1935; also J. Wayne Wrightstone, *Appraisal of Experimental High School Practices*, 194 pages, New York: Bureau of Publications, Teachers College, Columbia University, 1936.

¹⁴ Eugene R. Smith, Ralph W. Tyler and staff, *Appraising and Recording Student Progress*, 500 pages, New York: Harper and Brothers, 1933.

school marks, the discriminating value of the items is of major importance. With the exception of a few easy items at the beginning of such a test for the purpose of building morale in the pupils taking it, the items should show a percentage of successes increasing progressively from the poorest pupils to the best.

Only the simpler processes need concern the classroom teacher, for there is considerable doubt whether the elaborate techniques are enough better than the simpler ones to justify the additional labor involved.

One of the writers¹⁵ has devised a method of item analysis that is simple enough to be performed by a reasonably conscientious high school student. This procedure is explained fully in Appendix B, pages 436-453, where a typical classroom test is analyzed. The method can be outlined as follows:

- 1 Administer the test and score the papers, preferably putting a red X beside each incorrectly answered or omitted question.

- 2 On the basis of each student's total score, find the 27 per cent of the persons tested who scored highest (had the fewest X's). Call this the "high" group. Find the 27 per cent who scored lowest (had the most X's) and call this the "low" group. For instance, if there were 60 persons tested, the number in the high group would be $0.27 \times 60 = 16.2 = 16$. The number in the low group would also be 16. This would leave in the "middle" group $60 - (16 + 16) = 28$ papers to be put aside, for they are not needed in the item analysis.

- 3 Start with Item 1 on the test. How many persons in the low group missed it? How many persons in the high group? Subtract the number of persons in the high group who missed the item from the number in the low group who missed it.

- 4 Repeat Step 3 for every question in the test, each time determining the difference between the number of students in the low group who missed the item and the number in the high group who missed it.

- 5 List the differences in ascending order, beginning with the highest negative one and going down to the largest positive one, together with the item's number in the test. There should be as many differences as there are test questions. The largest positive differences (near the bottom of the list) indicate the most discriminating items, while the small positive differences and the negative differences suggest that those items are not discriminating properly and should therefore be looked over carefully for vagueness, improper keying, or unattractive options.

For the complete process, see Appendix B.

The chief value of item analysis to the teacher is that it helps make better

¹⁵ Julian C. Stanley. "A Simplified Item Analysis Procedure." *American Psychologist* 6: 749 July 1951. Abstract of paper read at the American Psychological Association convention in Chicago on September 1, 1951.

tests by pointing out unsuspected flaws in items that would otherwise probably continue to appear in these and later questions

I frequently, perhaps generally, it will be found that the trouble is in the *wording* of the item, the language being vague, ambiguous or positively misleading. In that case a rewording of the item may be all that is necessary. At times, however, the difficulty is more obscure and the item may have to be eliminated altogether. Lindquist found that "adequately" and "advisers" were equally difficult for eighth grade spellers but that the former discriminated in favor of the good spellers and the latter in favor of the poor spellers. Difficulty alone, therefore, is not a dependable measure of discrimination, for according to that criterion both items are equally good. Test experts have usually found however that the average difficulty of the items in a test is related to the adequacy of the test as a whole. The rule suggested for the construction of tests to discriminate best among all the members of a group is to make every item of 50 per cent difficulty when corrected for chance,¹⁶ so far as possible. This will mean that virtually all items of 0-15 per cent and 85-100 per cent difficulty when corrected for chance will be omitted from the revised form of the test unless they can be rewritten to make them closer to the 50 per cent difficulty level.

Tests designed primarily for *instructional* purposes however, may at times be made much easier with good results.

Judging the validity of standard tests. It is always desirable to examine with some care the content of a standard test before deciding to use it. Some of the earlier tests in particular contained serious errors. Upton¹⁷ called attention to some of these in arithmetic tests and Diamond¹⁸ found 318 errors in 3 303 items making up the content of sixteen widely used tests in biology and general science. Only one test was found to be entirely free from error. A study of five tests of English usage revealed that from 16 to 55 per cent of the items called wrong were actually acceptable according to standards published by the National Council of Teachers of English.¹⁹ Even when there are no errors the items used often stress the relatively unimportant aspects of the subject.

The test manual also should be examined, because it frequently gives data on the validity of the test. For example it should tell who made the test, how the items were chosen, what standardization and validation procedure was followed, and other pertinent information. If the author does

¹⁶ Correcting test scores for guessing is discussed on page 156. Items are corrected in a similar manner as explained in footnote 32 on page 160.

¹⁷ Clifford B. Upton. *The Influence of Standardized Tests on the Curriculum of Arithmetic*. *Teachers College Record* 26: 627-641 April 1925.

¹⁸ Leon N. Diamond. *Testing the Test-Makers*. *School Science and Mathematics* 32: 490-502 May 1932.

¹⁹ Karl W. Dykema. *On the Validity of Standardized Tests of English Usage*. *School and Society* 50: 767 December 9 1939.

not give such information to the prospective users it is safe to assume that the test is of doubtful validity, for it is evident that the author does not attach as much importance to the matter as is desirable.²⁰ While it is unnecessary to ascribe improper motives to test authors and publishers, most of whom are of a very high type, it is important to recognize that they are nevertheless human, and it is reasonable to make some allowance for a little overenthusiasm about the merits of their own progeny. Whenever available, therefore, the results reported in the professional literature by other users are likely to be especially valuable.

A systematic attempt to provide the information required by the test user as a basis for an intelligent choice of tests has been made by Buros.²¹ In a series of *Mental Measurements Yearbooks* he plans to make available critical evaluation of all recent tests by one or more competent persons independently. These publications will be found indispensable in selecting tests. The reviewer is instructed to make the reviews "frankly critical" and "to base the appraisals upon his own criteria as to what constitutes a good test." The reviewers are described as follows:

In selecting reviewers an effort was made to choose persons representing a wide variety of positions and viewpoints among actual and potential test users. As a result a very heterogeneous group of reviewers have cooperated in the preparation of this volume—classroom teachers, city school research workers, clinical psychologists, curriculum specialists, guidance specialists, personnel workers, psychologists, subject-matter specialists, and test technicians. It can be truly said that the reviewers represent no one group or school of thought, unless the reviewers are described as representing all test users—actual and potential—who are considered especially competent in their fields and who have the courage to speak frankly and honestly in appraising a standard test.

Ideally a reviewer of a standard test, such as a high school Latin test, ought to possess the qualifications of a curriculum and teaching specialist, and a test technician. Unfortunately all of these qualifications are rarely found in any one person. The average quality of the reviews is likely to be highest when the

²⁰ The American Psychological Association and other similar professional organizations have become concerned with the quality of test manuals and the methods of distributing tests. With regard to interest inventories, personality inventories, projective instruments and related clinical techniques and tests of aptitude or ability," see "Technical Recommendations for Psychological Tests and Diagnostic Techniques," *Psychological Bulletin*, 51: 1-38, March 1954.

A more general source is "Information Which Should Be Provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests," papers read by Herbert S. Conrad, Paul L. Dressel, and Laurence I. Shiffer, *Proceedings of the 1949 International Conference on Testing Problems*, pages 63-60, Princeton, New Jersey: Educational Testing Service, 1949.

²¹ Oscar K. Buros, *The Fourth Mental Measurements Yearbook*, Highland Park, New Jersey: Gryphon Press, 1953, 1189 pages. This yearbook covers the years 1948 through 1951 and therefore supplements rather than supplants the older yearbooks.

For bibliographies and discussions of tests during the period August 1, 1949, through July 31, 1952, see Frederick B. Davis (Editor), "Educational and Psychological Testing," *Journal of Educational Research*, 23: 1-110, February, 1953.

reviewers discuss only those points which they feel most competent to appraise, even though this practice frequently results in reviews which are not comprehensive.²²

The group judgment of even the most competent persons has certain limitations. There remains the troublesome fact that no one test is equally valid for all purposes, or for the same purpose in all situations. Furthermore, there is no way of knowing when a new test may appear with merits so outstanding as to render obsolete earlier tests that have hitherto been entitled to high comparative ratings. But, all things considered, the best available sources of information are probably the measurement specialists in reputable colleges and universities, of which there are several in most states. Their recommendations can usually be relied upon to be impartial and based upon a wider acquaintance with existing tests than the average teacher or school administrator is likely to have. But in the final analysis, when all the cards are on the table, the teacher or administrator must rely upon his own judgment. The necessary background for making such a judgment intelligently should be specifically provided for in the professional training of teachers. The data required for such judgments should be made available by the test publishers and by such publications as the *Mental Measurements Yearbooks*.

C. Reliability

Meaning of reliability. By reliability is meant the degree to which the test agrees with itself. To what extent can two or more forms of the test be relied upon to give the same results, or the same test to give the same results when repeated? If the scores on the test are stable under these conditions, the test is said to be reliable. In a word, *reliability means consistency*.²³

The terms *reliability* and *validity* are often confused, but there is a clear-cut distinction between them. Reliability, as such, has nothing to do with the truthfulness of the measurement, but is concerned only with its consistency, an entirely different thing. A homely illustration may help to clarify the distinction. A man returns from his vacation with a picturesque story of the fish he claims to have caught. As he meets friend after friend, there is always the same glowing account, even to the minutest detail. Now, in a statistical sense the story is *reliable*, for it is certainly *consistent*. Unfortunately, the fisherman's veracity is not thereby established, for consistency by itself gives no assurance of truthfulness or validity. In reality the story might be sheer fiction from beginning to end.

Importance of reliability. Shakespeare said: "Consistency, thou art a jewel," and he was right. But consistency is not the greatest jewel,

²² Oscar K. Buros, *The Nineteen Forty Mental Measurements Yearbook*, pages 12-13. Highland Park, N. J.: Gryphon Press, 1941.

²³ For a thorough discussion of this concept, see Robert L. Thorndike, "Reliability," in E. F. Lindquist (Editor), *Educational Measurement*, pages 560-620. Washington, D. C.: American Council on Education, 1951.

whether in a test or elsewhere. By itself consistency, or reliability, is a doubtful virtue, for a test as well as a person, might be consistently wrong but its absence is a sign of weakness. Although high reliability is no guarantee that the test is good, low reliability does indicate that it is poor. In the above illustration it should be noted that had marked discrepancies occurred in the fisherman's story from time to time, considerable doubt would have been cast upon his truthfulness. Validity is always the first quality to be sought in a test and, granted that reliability is a valuable auxiliary, *The ideal test tells the truth consistently*.

There can be little question that test makers have given too much attention, relatively, to determining the reliability of tests, and too little to establishing their validity. One reason for this, doubtless, is that the former is easier to determine. Much harm has resulted, however, when uncritical users have naively assumed that reliability insures validity, a view which is wholly erroneous.

Methods of determining reliability. The term *reliability* is purely a statistical concept. Contrary to what was found in the case of curricular validity, little can be told about the reliability of a test from examining the test blank itself. It is of course true that if a test can be objectively scored it is more likely to be reliable than if the scoring is subjective, but the degree of reliability cannot be determined by that fact. It is also true that a long test has a greater likelihood of being reliable than a short test but there are many exceptions. In the last analysis, however, *somebody must try the test out to determine its reliability*. Usually the author of the test does this and reports the results in the test manual. If such is not the case one has a right to be suspicious of the merits of the test.

Method with two test forms. Three rather distinct techniques are used to establish the reliability of a test. The method commonly used by makers of standard tests is to prepare two or more parallel forms of the test, and then to give these equivalent forms of the test to a large number of pupils usually with only a short interval between the tests. The test is said to be reliable if there is close agreement between the scores on the two forms that is if the pupils who made high scores on the first test also make high scores on the second if those who made low scores on the first test again make low scores on the second and so on for all ranks in between. If the agreement is perfect as is most unlikely the correlation is 1.00. On the other hand, if there is no consistent relationship the coefficient of reliability is .00. It will be recalled that validity is also expressed as a coefficient of correlation whose maximum value is 1.00, and whose minimum value is .00. But in the case of statistical validity the agreement is with an external criterion whereas in the case of reliability the agreement is with an internal criterion of some kind. In the above illustration this internal criterion is another form of the same test, which presumably measures the same functions as the first test.

Methods with one test form. When only one form of a test is available, probably its reliability can still be determined. One procedure is to repeat the test at a later time and to determine the extent of agreement by computing the coefficient of correlation between the two series of scores. Another procedure is to give the test once only and then to record two scores for each paper, one for each half. The test is split into halves matched for content and difficulty. When the two series of scores are obtained the coefficient of correlation between them is computed. This is the reliability of the half test. The reliability of the whole test is then estimated by the use of a special formula.²⁴ A similar formula also makes it possible to estimate the probable reliability of the test when increased to any required length, assuming that the items added are of the same type and quality as those in the original test.

Both methods for obtaining the reliability of a single form of the test have been severely criticized and as stoutly defended. The test-retest method has certain serious limitations. If the test is long, to avoid fatigue and boredom some time must elapse between the two trials. In the case of achievement tests, particularly, this delay is likely to introduce other variables. The pupils may discuss the test between trials, do extra study, or do other things that may effect a change in the status of their knowledge. In addition to this, their physical and mental conditions fluctuate from day to day, even from hour to hour. For example, Ashbaugh²⁵ found variability in one fourth of the pupils who were given the same spelling test under highly constant conditions three times within fifteen minutes. One would appreciate the difficulty in determining the reliability of a certain type of thermometer by checking the readings made at one hour against those made later in the day. Guilford thinks that "it is safe to say that the average test scale of mental ability is fully as reliable, or probably more so, than the average clinical test in medicine, such as the test of blood pressure or the basal metabolism test, whose reliability ranges from about 60 to 90."²⁶ But there is also the contrary tendency in human beings for errors made the first time to persist and to be repeated at later times. In an extreme situation, where the pupils memorized the first series of answers, the apparent reliability of the test would be perfect. Indeed, this tendency to echo the original responses appears to be strong, for test-retest coefficients are usually higher than those arrived at by correlating halves or equivalent forms of the test. Because the correlation of the half tests eliminates, or

²⁴ This is the simplest form of the Spearman-Brown 'prophecy' or 'step-up' formula which appears in many test manuals. The reliability of the whole test equals twice the r between the half test scores divided by $(1 + \text{the } r \text{ between half test scores})$ being

written in symbols as $r_w = \frac{2r_{12}}{1 + r_{12}}$

²⁵ Ernest J. Ashbaugh, Variability of Children in Spelling, *School and Society*

q 93 38 January 18 1919

²⁶ J. P. Guilford, Intelligence Tests, *Education* 58 228 May 1935

at any rate greatly reduces, memory carry over, it is recommended by some writers

The split-half procedure, involving as it does the Spearman-Brown formula, is based upon certain assumptions. The half-tests should be of equal variability, and the items in one half must be of the same quality as those in the other half. It must be emphasized that the formula requires the use of matched halves of the test, not just any halves.²⁷

Kuder and Richardson²⁸ have devised several methods of obtaining a reliability coefficient which make it unnecessary to split the test into halves or calculate a coefficient of correlation. Unfortunately, their most usable formula, called KR No. 20, is very laborious for the classroom teacher to compute.

As Cronbach²⁹ aptly points out, the comparable-forms, split-half, and test-retest reliability coefficients get at different aspects of reliability. The first is a "coefficient of equivalence and stability," the second a "coefficient of equivalence" only, and the third a "coefficient of stability."

A strong word of caution is needed. *Neither the split half method nor the various Kuder-Richardson formulas are applicable to "speeded" tests*, for they overestimate their reliability. A test is speeded when many of the examinees would have made better scores had they been given more time. If nearly all persons did about as well in the time allowed as they could have done in a longer period, then the measuring instrument is a "power" test. In practice, most timed tests involve both speed and power. Watch for the rather frequent split-half or KR reliability coefficients reported in test manuals for speeded tests, they are deceptively high.

One of the writers has published a simplified computational technique for securing split-half reliability coefficients from unspeeded tests.³⁰ Another shortened procedure is illustrated in Appendix B on pages 452-453.

The interpretation of test reliability. What standard shall a test meet in order to be considered satisfactory from the standpoint of reliability? No simple answer to this question is possible. It depends, for one thing, upon the fineness of discrimination required. Kelley³¹ has suggested

²⁷ However, Clark has shown that the variation of split-half r 's from sample to sample is much greater than the variation of such r 's within a sample due to different methods of splitting, provided that the method of splitting is longitudinal—puts items through out the test in each half rather than having one half consist of the first items and the other half of the last. See Edward L. Clark, 'Methods of Splitting vs. Samples as Sources of Instability in Test Reliability Coefficients,' *Harvard Educational Review* 19: 178-182, May, 1949.

²⁸ G. F. Kuder and M. W. Richardson, 'The Theory of the Estimation of Test Reliability,' *Psychometrika* 2: 151-160, September 1937.

²⁹ Lee J. Cronbach, *Essentials of Psychological Testing*, pages 65-73, New York: Harper & Brothers, 1949.

³⁰ Julian C. Stanley, 'A Simplified Method for Estimating the Split-Half Reliability Coefficient of a Test,' *Harvard Educational Review*, 21: 221-224, Fall 1951.

³¹ Truman Lee Kelley, *op. cit.* pages 28-29.

the following minimal requirements for the reliability coefficients of a single school grade:

- .50 for determining the status of a group in some subject or group of subjects
- .90 for differentiating the achievement of a group in two or more scholastic lines
- .91 for differentiating the status of individuals in the same subject or group of subjects
- .98 for differentiating individuals in two or more scholastic lines

The interpretation is also beset by many other difficulties. The coefficients not only reflect somewhat the methods employed in their computation, but also the variability of the groups, the interval between tests, and other factors. For example, a test of average difficulty for a typical group may be much less reliable when used with a markedly inferior or markedly superior group.

In view of the fact that measures of reliability, no matter how arrived at, are influenced by factors other than the form and content of the test itself, it would appear that the value of such measures has been overemphasized.²² The same energy devoted to improving the validity of the test would bring better returns. It is not likely that the average teacher will find it profitable to compute reliability coefficients for ordinary class tests, although it may sometimes be worth while to do so for final examinations.

Objectivity and reliability. By objectivity in a measuring instrument is meant the degree to which equally competent users get the same results. Ordinary measures of height and weight, for example, are objective, while estimates of beauty and integrity are subjective. The distinction between objective measurement and subjective measurement is implied in the question: "Do married men *really* live longer than single men, or does it just *seem* longer?" As a rule, objectivity is very closely associated with reliability. For this reason standard tests are usually more reliable than rating scales. As a matter of fact, great impetus was imparted to the objective test movement by the discovery that the major cause of the notorious unreliability of the ordinary school examination was its subjectivity of marking. The emphasis on objectivity has since gone so far, however, that many educational workers seem to regard "objectivity" as synonymous with "scientific method." To such persons, any element of subjectivity in a study renders it hopelessly unscientific. It may be well, therefore, to look carefully at this all-important matter of objectivity.

To discover at the outset that there is no such thing as a wholly objective measure may be something of a shock. The plain fact is that objectivity is always relative, never absolute. The measurements obtained by a yard-

²² Some writers would abandon altogether the "blanket term reliability" in favor of more specific estimates of absolute and relative accuracy of measurement. Robert W. B. Jackson and George A. Ferguson, page 25 in *Studies on the Reliability of Tests*, Bulletin No. 12 of the Department of Educational Research, University of Toronto, 371 Bloor Street West, Toronto 5, Ontario, Canada, 1941.

stick, for example, are only relatively objective, for one would hardly expect a dozen different persons to get absolutely the same results in measuring the length of the playground. They would probably agree to the nearest foot, and possibly to the nearest inch, but they would usually disagree markedly, if the results were expressed in some such small unit as hundredths of an inch. And, of course, such units as inch, foot, and yard are not *natural* units, like day and year, but units set up by human judgment.

Brownell points out that there are always many subjective factors involved even when the test used is of the so-called objective type. He says:²³

Well, first of all, in the practical circumstances of teaching one *decides to give* a test. The decision is surely not based upon purely objective considerations. Second one determines whether to *make* a test or to *buy* one. Third, one makes up one's mind regarding the *kind* of test—whether it is to be of the traditional type, of the newer types, or a combination—judgment again. Fourth one settles upon the *scope* of the test—judgment once more. Fifth, one selects the *items* to be included—little objectivity here. Sixth, one chooses the *form* to be employed—true-false, multiple choice or what not—again little objectivity. Seventh, one *frames the items* as carefully as one can—and once more has only his judgment for guidance. Eighth, one prepares a *key* by listing the correct answers—a judgment which may not be acceptable to other teachers even of the same subject. Ninth, through opinion one defines the conditions of *administering* the test. Tenth, one *scores* the papers—at last objectivity. But, eleventh, one *assigns marks*—another increment of judgment, and a big one.

Brownell protests against what he regards as the overemphasis on objectivity, which he thinks has unnecessarily lessened the depth and narrowed the range of measurement. A safe position would appear to be to *try to make measurement as objective as possible without sacrificing validity*. It must be remembered that the latter is always more important. It is never going to be possible or desirable to eliminate certain basic assumptions underlying all attempts at evaluation. Undoubtedly at times, however, we have made *assumptions* in measurement when we should have had *evidence*. Many test makers, for example, have assumed that one problem of a type is sufficient for diagnosis in arithmetic. When the matter was actually subjected to experimental analysis in two studies,²⁴ both found that one problem of a type is likely to be both unreliable and invalid, owing largely to chance, and that at least three problems of each type must be included for satisfactory individual diagnosis. Another assumption which did not check with the evidence was that objectivity of scoring guarantees accuracy of scoring. Several studies have demonstrated the fact that scorers

²³ William A. Brownell, "The Use of Objective Measures in Evaluating Instruction," *Educational Method* 13: 401-408, May-June 1934.

²⁴ Leo J. Brueckner and Mary Ellwell, "Reliability of Diagnosis of Error in Multiplication of Fractions," *Journal of Educational Research* 26: 175-185, November 1932; Foster F. Grounckle, "Reliability of Diagnosis of Certain Types of Error in Long Division with a One-Figure Divisor," *Journal of Experimental Education* 4: 7-16, September 1933.

of standardized tests must be *taught* and not merely *told* how to do it³⁵ A serious effort should of course be made to eliminate all needless types of subjectivity A guess is usually a poor substitute for actual knowledge

D Usability

Meaning of usability There is quite general agreement among authorities in measurement that the two most important characteristics of a measuring instrument are validity and reliability Both have to do with the theoretical accuracy with which the instrument measures However there are certain other considerations of a practical character which must be taken into account In the judgment of the writers all of these may be conveniently designated by the single term *usability* By this is meant the degree to which the test or other instrument can be successfully employed by classroom teachers and school administrators without an undue expenditure of time and energy—in a word *usability means practicability* A measuring instrument must not only be valid and reliable but also usable This viewpoint is well expressed in *The Methodology of Educational Research*³⁶

But we must always temperize ideals with practical considerations Perhaps an ideal instrument would be so cumbersome and expensive of effort and time that its use would not be warranted

Whether or not a test is usable by average teachers in service and other persons whose technical training in measurement has been limited depends upon several factors, of which the following are probably the most important

- 1 Ease of administration
- 2 Ease of scoring
- 3 Ease of interpretation and application
- 4 Low cost
- 5 Proper mechanical make-up

Each of these factors will now receive brief consideration

Ease of administration Group tests, as a rule are much easier to administer than individual tests The Stanford Binet is a good example of a test whose validity and reliability are high but whose usability is low largely because of complicated instructions for giving and scoring Special training in a college course for one semester is usually suggested as the minimum required for mastery of these instructions Even then the test makes heavy demands upon the examiner's time

There are of course two types of instructions for a test One has to do

³⁵ For helpful suggestions see Arthur E Traxler *Administering and Scoring the Objective Test* in E F Lindquist (Editor) *Educational Measurement* pages 329-416 Washington D C American Council on Education 1951

³⁶ Carter V Good A S Barr and Douglas E Scates *The Methodology of Educational Research* page 439 New York D Appleton Century Company 1936

with directions to the examiner, and the other has to do with directions to the pupil or pupils. But, in general, the requirements are the same for both. The motto of a certain news weekly indicates what is required: the directions should be "clear, curt, complete." Whether or not examples, fore-exercises, and the like are necessary will depend mainly upon the age and experience of the group being examined. Whether or not a group test is easy to administer depends to a considerable extent upon the completeness of the manual. Some tests have no time limits, many have generous time limits, while still others are broken up into intervals as short as 3, 5, 8, 10, or 15 seconds. These short intervals are difficult to observe with a stop watch and well-nigh impossible without it. On the other hand, tests of the so-called self-administering type involve only one short set of directions for the entire test. Most tests, however, are broken up into separate sections, each of which has its own directions and time limit. In determining how difficult a test is going to be to administer, a careful examination must be made both of the manual and of the test blank itself.

Ease of scoring. The ease of scoring a test depends primarily upon three things: objectivity, adequate keys, and full scoring directions. The better standard tests rank high on all three counts. Scoring is also facilitated when the pupil has been instructed to record his answers in a straight column rather than irregularly over the page, and in the form of a numeral or single word rather than a phrase or longer statement. As a rule, all acceptable answers should appear on the key. With the exception of scales in which score values of unequal weight are required, each correct item should count the same as any other correct item in the test. The unequal weighting of items, so common in the earlier tests, has been found to add to the difficulty of scoring without a corresponding increase in validity or reliability.

Three ways to speed up the scoring of objective tests are (1) hand-scorable separate answer sheets, (2) machine-scorable separate answer sheets, and (3) self-scoring answer sheets. Traxler says:

The growing tendency to employ separate answer sheets is perhaps the most pronounced single trend in objective testing. It unquestionably has a restrictive effect upon test construction for it tends to force test items into a single-response pattern—the multiple choice question. The use of other types of questions is not precluded, however, if test makers will use skill and ingenuity in setting up their answer sheets. More imagination and care than are typically employed in devising a test need to be used when setting up separate answer forms.

There is not, at present, enough experimental evidence to warrant a definite statement concerning the effect of separate answer sheets upon the validity of the test results. It is appropriate to urge, however, that test authors have an inescapable obligation to find by means of research an answer to this question before they "go overboard" completely for answer sheet procedures.

The main justification for the use of separate answer sheets is that they are ac-

¹ Arthur I. Traxler, *op. cit.* page 41.

expensive, that they save time, and that they bring objective tests within the reach of hundreds of schools that could otherwise not use these tests at all. Through the continued availability of these inexpensive instruments, objective testing may eventually be brought to all schools in the United States.

Concerning machine scoring Traxler is less optimistic, pointing out that even for very large testing programs "Machine scoring may not be faster or cheaper than manual scoring of a prescribed uniform program where all manual scoring procedures can be highly routinized."³³ Furthermore, unless carefully supervised, machine scoring may be less accurate than independently checked hand scoring. The chief advantage of machine scoring seems to be in those situations where the machine is kept going all day long and is operated by a full-time highly skilled person.

Many self-scoring devices are available.³⁴ Typical of these are "hat-pin" punching methods utilized in the Kuder Preference Record³⁵ and the Ohio State University Psychological Test,³⁶ the concealed carbon of the Clapp-Young self-marking tests³⁷ and Scoreze³⁸ and the "punchboard" procedure of the Science Research Associates self-scorer,³⁹ where the student punches until he finally obtains the correct answer. All of these probably have considerable merit as instructional and motivational devices, but if scores are desired for evaluational purposes they can at the present time usually be secured more easily and dependably by other methods, except perhaps for the Kuder Preference Record.

Ease of interpretation and application. Whether or not the results of a test are easy to interpret and apply depends primarily upon the adequacy of the manual accompanying the test. In the first place, the manual should contain complete norms to facilitate interpretation. Whenever possible all derived scores should be capable of being read directly from tables of norms without the necessity of computation. The norms should, as a rule, be based both on age and on grade, and, in the case of high school achievement tests, on the length of time the subject has been studied. It is also desirable that achievement tests be provided with separate norms for urban and rural pupils, and for pupils of various degrees of mentality. Up to the present time very few tests are adequately provided with norms for interpretation. Where the primary emphasis is upon diagnosis and other instructional values of tests, this loss is not very great. In any event, it will usually be necessary to rely heavily upon local norms.⁴⁰

³³ Arthur T. Traxler *ibid.* page 408.

³⁴ Sidney L. Pressey, 'Development and Appraisal of Devices Providing Immediate Automatic Scoring of Objective Tests and Concomitant Self Instruction' *Journal of Psychology* 29 417-447 April 1950.

³⁵ Published by Science Research Associates.

³⁶ Published by Ohio College Association, Ohio State University, Columbus 10, Ohio.

³⁷ Published by Houghton Mifflin Company.

³⁸ Published by California Test Bureau.

³⁹ Published by Science Research Associates.

⁴⁰ A fuller discussion of norms will appear in Chapter 10.

Several of the better manuals give specific suggestions regarding the use to be made of the test results.⁴⁶ Supplying some suggestions as to results is a valuable service for which it is hoped test publishers in the future will accept more responsibility. For many uses it is necessary to have at least two forms of the test equated both as to content and as to difficulty throughout the full range of scores, and not for averages only. Few tests meet this requirement fully.

Cost. With the exception of certain laboratory apparatus and equipment for measuring special abilities and disabilities, testing materials are usually not very expensive. Few achievement tests covering a single subject, or group tests of general intelligence, cost more than ten cents each. Batteries covering several subjects when printed as a single booklet usually cost from ten to fifteen cents. For a comprehensive testing program the general battery will cost less than separate tests covering the same subjects. Cost is a practical consideration in most school systems, and there is no point in paying more for tests than necessary.

While it may be true in general, as commonly held, that in the long run one gets about what he pays for, there are too many exceptions to make it a safe rule. In statistical terminology the correlation between the cost of a test and its worth is positive, but too low for accurate prediction. Here, as elsewhere, the customer should be wary, lest he not get his money's worth. In a test, as in an automobile, the quality is often not evident on the surface. The prospective purchaser should not make cost a primary consideration, for good tests are often no more expensive than poor ones. Fortunately, therefore, relative cost can be considered a minor matter, as a rule, and the choice of the test can rest, as it should, upon its validity and reliability for the purpose it is to serve. It must be remembered that one test may be cheap enough at fifteen cents and another too costly at five. After all, the careful purchaser is more concerned with what he gets for his money than with what he has to pay.

Mechanical make-up of the test. Tests issued by the larger publishers are almost always printed in clear type of a size appropriate to the grade level for which they are intended. But there are some exceptions. One of the leading publishers issued a test in which the key word in each sentence was supposed to be in bold faced type, but in many cases the poor quality type did not clearly indicate which word was intended. On a timed test such as this one, a handicap was imposed upon all pupils except those with the keenest vision. In the lower grades careful attention should be given to the quality of pictures and illustrations used. In the earlier days it was

⁴⁶ A good example is Gertrude H. Hildreth and Harold H. Bixler, *Manual for Interpreting the Metropolitan Achievement Tests* (Yonkers-on Hudson, New York: World Book Company, 1915, 122 pages). This manual is not given free with test orders, however its price is low.

common to have the instructions to the examiner appear on the test booklet in the hands of the pupil. This practice not only meant a needless cost in paper and printing, but was possibly a source of confusion to the pupil.

Commercial publishers of tests have not as yet given sufficient attention to devising tests which will reduce to the minimum their cost in time and money. There appears to be no valid educational reason why most tests should not be designed with separate answer sheets, a practice which not only eliminates the economic waste of using the test one time only and then discarding it, but which also greatly facilitates the scoring and makes the pupil's test profile available in convenient form for use and filing. Little is gained, however, when the answer sheet itself is sold at prices almost as high as the test itself, as is now often the case. The customer usually gets what he wants when he wants it badly enough and makes his wishes known. To make convenient, inexpensive tests feasible, moreover, the demand must be sufficiently increased so that the additional volume sold compensates for reduced profit per unit.

Summary. What, then, are the earmarks of a good measuring instrument? In brief, a good test or other measuring instrument possesses three outstanding qualities: validity, reliability, and usability. In other words, *a good test measures what it claims to, consistently, and with a minimum expenditure of time, energy and money.* But always the first consideration is validity. The test must not only measure what it purports to, but in the case of achievement tests, it should purport to measure the really important outcomes of the educational process. No achievement test that fails to do this can be considered a satisfactory measuring instrument, whether made by the classroom teacher or purchased from a publisher of standard tests.

E. Some Generalizations Regarding the Problem of Measurement

The role of measurement in science in general and in education in particular was set forth in the first chapter, the historical development of measurement in education was traced in the second chapter, quantitative concepts were discussed in the third chapter, and the characteristics of a satisfactory measuring instrument were described in the present chapter. In the light of these discussions a few important generalizations will now be attempted.

1 *Some kind of measurement or evaluation is inevitable in education.* This generalization is amply supported by the history of every recognized science, and of education itself, regardless of whether it is to be classified as a full-fledged science or not.

2 *All measurement is subject to error.* This is true of the so-called "exact sciences", and to a greater degree it is true of the less exact or newer social

sciences, such as psychology and education Westaway, for example, thinking mainly of physics and chemistry, concludes "We may, in fact, look upon the existence of error in all measurements as the normal state of things"⁴⁷ Kelley speaks of the "ubiquitous probable error"⁴⁸ in psychology and education These errors can be reduced but never wholly eliminated

3 *These errors of measurement are due in part to the imperfection in the measuring instruments available* There are no perfect measuring instruments, even in the physical sciences Westaway, for example, says that "even the very best of the instruments with which we perform our measurements are imperfect"⁴⁹ This is true of the fundamental units of measurement in the physical sciences, as well as of the biological and social sciences No astronomer knows precisely the velocity of light, and yet the light year is the yardstick of celestial measurement, no chemist knows the precise value of a single atomic weight, and yet it is the basic unit in chemical analysis In psychology and education these imperfections are an even more potent source of error than in the older sciences However, it must be remembered that the tools of measurement are much better than they used to be

4 *The limitations of the methods used are a still more important source of error in measurement* Again this difficulty is true of the physical sciences as well as of the social sciences For example, Max Planck says that in physics "every measurement, however exact, inevitably involves certain errors of observation"⁵⁰ These errors are due partly to sensory and temperamental defects, and partly to lack of skill in the observer But a still more troublesome source of error is the tendency for the act of observation to interfere with the phenomena being observed in measurement Heisenberg, for example, noted that the "measurement of an electron's velocity is inaccurate in proportion as the measurement of its position in space is accurate, and vice versa,"⁵¹ owing to the disturbing influence of the light rays falling on it in the act of measurement From this discovery resulted the famous "uncertainty principle" or the "principle of indeterminacy," which has profoundly influenced modern physics "As a matter of fact every measurement," says Planck, "whatever the method of its employment, invariably interferes more or less with the event to be measured"⁵² But this interference is so slight as to be only of theoretical interest to the laboratory physicist engaged in the study of aggregates of elements instead of individual electrons And the ordinary Newtonian principles of chem-

⁴⁷ F W Westaway, *Scientific Method Its Philosophical Basis and Its Modes of Application* pages 289-290 New York Hillman Curl Inc, 1937

⁴⁸ Truman Lee Kelley *op cit* page 19

⁴⁹ F W Westaway *op cit* page 286

⁵⁰ Max Planck *The Philosophy of Physics* page 24 New York W W Norton & Company Inc 1936

⁵¹ *Ibid* page 62

⁵² *Ibid* page 69

istry and physics still operate in the usual way in such practical realms as engineering and medicine.⁵²

But the disturbing effect of the measurement process is more serious in education. The personality of the examiner, as well as the testing materials, is always part of the test situation. This is recognized in giving individual tests, where a proper *rapproch* between examiner and subject is regarded as essential to a successful examination.⁵⁴ But here, even with skilled examiners, the factor is rarely eliminated altogether, for it has been found that the IQ remains more stable when the same examiner gives all the tests. In all experiments, whether involving the use of individual or group tests, the subjects are not purely naïve and receptive creatures but are actuated by motives of pride, desire to please or make a good impression on the examiner, and the like. In other words, the examiner or experimenter is an important part of the situation, and it is doubtful whether standardized instructions can ever reduce this part to the point at which it is negligible. The factor is especially important in personality measurement and in the evaluation of social behavior. Certainly one would hardly expect to get as normal reactions of love-making in the psychological laboratory during the day as he would if he were concealed in a tree beside a bench in the park during the evening.

5 Teachers and school administrators must not only understand and appreciate the functions of measurement in education, but they must realize more fully the limitations of present measuring instruments. In the present state of measurement two erroneous attitudes are sometimes found. The first is that held by certain over-enthusiastic supporters of measurement, who make unreasonable claims for existing measuring instruments, and who gloss over or refuse to recognize the imperfections that exist.⁵⁵ This attitude is not unlike that of the adolescent in his first love affair, where, indeed, if love is not actually blind, it deliberately closes its eyes, and in any event the result is the same. This point of view is unfortunate and unintelligent, for it stands in the way of progress toward needed improvements, making such unwarranted claims is the surest way to discredit the movement with thoughtful people.⁵⁶ Fortunately, this attitude appears to be on the decline.⁵⁷

⁵² For a stimulating discussion of the practical implications of this uncertainty principle by a distinguished American chemist see Irving Langmuir, *Science, Common Sense, and Deceit*, *Science News Letter*, 43: 3-4, 12-15 January 2 1943.

⁵⁴ Elmer L. Sacks, "Intelligence Scores as a Function of Experimentally Established Social Relationships Between Child and Examiner," *Journal of Abnormal and Social Psychology*, 47: 354-358, April Supplement 1952.

⁵⁵ For a suggestive analytical discussion of the problem see Douglas E. Scates, "Differences Between Measurement Criteria of Pure Scientists and of Classroom Teachers," *Journal of Educational Research* 37: 1-13 September 1943.

⁵⁶ Ross recalls having heard a gray haired southern educator say, regarding intelligence tests soon after World War I: "The worst enemies of any new cause are its darn fool friends!"

⁵⁷ Picturesque pleas for sanity in using tests by two pioneers in test construction are

But a second and equally erroneous attitude goes to the opposite extreme. It characterizes those who are as blind to the virtues of existing measuring instruments as the first group are to their limitations, and who refuse to have anything at all to do with tests and examinations until all defects are forever removed. This attitude is as unreasonable as that of the farmer who has postponed buying an automobile "till them blamed things is perfected," and who has in the meantime worn out a great deal of shoe leather without seeing much of the world either.

Then there is the third attitude—that of the practical person who has learned through experience not to expect perfection. Moreover, he has found that excellent work can often be turned out with imperfect tools, if only they are used with sufficient skill. He has also discovered that greater skill is called for than if the instruments were perfect, and he sets out deliberately to attain the skill needed. He realizes that the very existence of these imperfections imposes a special obligation upon the user to seek to understand as fully as possible their nature in order to get desired results in spite of them. Furthermore, he makes a conscious effort in interpreting and using test results in order to take into account the existence of errors. In other words, he takes the very common-sense point of view that the proper thing to be done under the circumstances is to make the best possible use of such tools as exist, while waiting for better ones to be developed.

SELECTED REFERENCES FOR FURTHER READING

- Bennett, George K., Seashore, Harold G., and Wesman, Alexander G., *Differential Aptitude Tests Manual* (Second Edition) New York: The Psychological Corporation, 1952. Chapter 4, "Validity," and Chapter 5, "Reliability."
- Brownell, William A. (Chairman), "The Measurement of Understanding," *Forty-Fifth Yearbook of the National Society for the Study of Education, Part I* Chicago: University of Chicago Press, 1946. 338 pages.
- Davis, Frederick B., "Item Analysis in Relation to Educational and Psychological Testing," *Psychological Bulletin*, 49: 97-121, March, 1952.
- Dressel, Paul L., "Problems of Evaluation in General Education," *Proceedings of the 1951 Invitational Conference on Testing Problems*, pages 45-57. Princeton, New Jersey: Educational Testing Service, 1952.
- DuBois, Philip H., "Achievement Tests in Personnel Selection," *American Journal of Public Health* 41: 567-572, May, 1951.
- Fan Chung Teh, *Item Analysis Table*. Princeton, New Jersey: Educational Testing Service, 1953. 32 pages.
- Flanagan, John C. (Chairman), "Constructing Examinations So That They Will Be Valid Measures of Important Functions," *Proceedings of the 1948 Invitational Conference on Testing Problems*, pages 13-42. Princeton, New Jersey: Educational Testing Service, 1949. Papers by Oscar K. Buros, Max D. Engelhart, Warren G. Findley, Harold Gulliksen, Charles R. Langmuir, and Philip J. Rulon.

made in S. A. Courtis, "Let's Stop This Worship of Tests and Scales," *Nation's Schools* 31: 16-17, March 1913, and in Guy M. Wilson's "Some Subversive Activities of the Test Expert," *Educational Method* 21: 312-343, April 1912.

- Freeman, Frank S, *Theory and Practice of Psychological Testing* New York Henry Holt and Company, 1950 Chapter 1, "Basic Principles Theoretical and Clinical "
- Guilford, Joy P, *Fundamental Statistics in Psychology and Education* (Second Edition) New York McGraw-Hill Book Company, 1950 Chapter 17, "Reliability of Measurements," and Chapter 18, "Validity of Measurements "
- Lindquist L F (Editor) *Educational Measurement* Washington D C American Council on Education, 1951 Chapter 5 Preliminary Considerations in Objective Test Construction," by L F Lindquist Chapter 9, "Item Selection Techniques," by Frederick B Davis, Chapter 13, "The Essay Type of Examination," by John M Stulmaker, Chapter 14 "The Fundamental Nature of Measurement," by Irving Lorge, Chapter 15, "Reliability," by Robert L Thorndike and Chapter 16, "Validity," by Edward E Cureton
- Lindquist, L F, "Some Criteria of an Effective High School Testing Program," pages 17-33 in Arthur E Traxler (Editor), "Measurement and Evaluation in the Improvement of Education," *American Council on Education Studies, Series I*, No 46, Vol XV, April, 1951 Washington, D C American Council on Education
- Odell, C W, *How to Improve Classroom Testing* Dubuque Iowa Wilham C Brown Company, 1953 Chapter I, "Introduction " and Chapter II, "Objectives "
- Thorndike, Robert L, *Personnel Selection Test and Measurement Techniques* New York John Wiley & Sons, 1949 Chapters 4, 5, and 8 "The Estimation of Test Reliability," "The Estimation of Test Validity Criteria of Proficiency," and "Item Analysis and Selection of Items "
- Thorndike, Robert L (Chairman), "Criteria for the Evaluation of Achievement Tests," *Proceedings of the 1950 Invitational Conference on Testing Problems*, pages 73-112 Princeton, New Jersey Educational Testing Service, 1951 Papers by John B Carroll, Frederick B Davis, Harold Gulliksen, and Joseph J Schwab
- Thorndike, Robert L, "Tests as Research Instruments," *Review of Educational Research*, 21 450-462, December, 1951
- Travers, Robert M W, *How to Make Achievement Tests* New York The Odyssey Press, 1950 Chapter 1, "Introduction "
- Traxler, Arthur E, Jacobs, Robert, Selover, Margaret, and Townsend, Agatha, *Introduction to Testing and the Use of Test Results in Public Schools* New York Harper & Brothers, 1953 Chapter 1, "A Point of Departure " and Chapter 2, "What Do Tests Contribute to Understanding the Individual Pupil?"
- Vernon, Philip E, *The Structure of Human Abilities* New York John Wiley & Sons 1950 Chapter IV, "Analyses of Educational Attainments "
- Weitzman, Elhs, and McNamara, Walter J, *Constructing Classroom Examinations —a Guide for Teachers* Chicago Science Research Associates, 1949 Chapter 1, "Basic Aspects of Achievement Tests "

PART II

The Construction of Teacher-Made Tests

5

General Principles of Test Construction

Importance of the problem. There are at least three reasons why the development of proficiency in constructing informal teacher made tests is important. In the first place, the vast majority of tests in use by classroom teachers are of this type.¹ In the second place, both essay examinations made and marked by untrained teachers and objective tests used by ordinary classroom teachers may produce highly unsatisfactory results. The extensive literature on essay examinations, briefly summarized in Chapter 2, has repeatedly demonstrated this fact. Amateurs may at times do even worse with objective tests than with essay examinations. Incredible as it may be, it does seem possible to make objective tests of lower reliability than essay examinations.² In the third place, both logical considerations and statistical analyses indicate that skillfully prepared informal tests are as reliable and as valid as some standardized tests. In fact where the teaching conditions are unusual, or where the subject matter is not thoroughly stabilized, as in civics and modern history such tests may be even more valid. A state-wide survey of high school achievement conducted in Tennessee,³ for example, showed that only 56 per cent of the questions in the standardized social studies test then in use could be answered from the state-adopted textbook.

This chapter will consider the general principles of constructing informal

¹Iven H. Hensley and Robert A. Davis. What High School Teachers Think and Do about Their Examinations. *Educational Administration and Supervision* 38: 219-228 April 1952.

²John M. Stalnaker. The Essay Type of Examination. Chapter 13 in E. F. Lindquist (Editor) *Educational Measurement*. Washington, D. C.: American Council on Education, 1951.

³Joe E. Aynt. *Report of the Tennessee State Testing Program*, page 83. Nashville: State Department of Education, 1946.

teacher made tests grouped under the following four headings, which indicate roughly the steps or stages in the process

- 1 Planning the test
- 2 Preparing the test
- 3 Trying out the test
- 4 Evaluating the test

A Planning the Test

It should be recognized at the outset that the construction of satisfactory measuring instruments is one of the most difficult duties the teacher has to perform. Good tests do not just happen. Nor are they the result of a few moments of high inspiration or exaltation. On the contrary, the process is calm, deliberate, and time-consuming. Perhaps the best that can be hoped for under existing conditions is that the teacher prepare reasonably comprehensive and adequate informal tests in one subject each year. Best results will usually be obtained from cooperative effort. The procedure employed in developing the Cooperative Achievement Tests, outlined on page 112, is a good illustration. Another example is the cooperative plan used by 19 colleges in constructing examinations.⁴ A third illustration is the procedure followed by the Evaluation Staff of the Eight-Year Study sponsored by the Progressive Education Association.⁵ The six major steps in the process as set forth in detail by Smith and Tyler⁶ may be summarized briefly as follows:

- 1 The faculty of each school was asked to formulate a careful statement of its educational objectives.
- 2 Statements from these thirty schools were classified by the Evaluation Staff into ten major types of objectives.
- 3 Each type of objective was then defined in terms of expected pupil behavior.
- 4 Situations were suggested in which pupils could be expected to show the particular kind of behavior.
- 5 The more promising methods of obtaining evidence regarding each type of objective were then selected from existing techniques or devised by the staff, and subjected to experimental trial.
- 6 The methods which made the best showing in this preliminary trial were further developed and improved.
- 7 Means were devised for the interpretation and effective use of the various instruments of evaluation.

It is recognized that the process just described is too elaborate for the ordinary school or for the individual teacher working on his own. However,

⁴ Paul L. Dressel, *Problems of Evaluation in General Education*. *Proceedings of the 1st Institutional Conference on Testing Problems*, pages 45-57. Princeton, New Jersey: Educational Testing Service, 1932.

⁵ Published in five volumes by Harper & Brothers, New York, 1942, under the general title *Adventure in American Education*.

⁶ Eugene R. Smith, Ralph W. Tyler, and Evaluation Staff, *Appraising and Recording Student Progress*, Chapter I. New York: Harper & Brothers, 1942.

it cannot be emphasized too strongly that the actual process of test construction must be preceded by careful planning if the test is to succeed. The test will be no better than the quality of the thinking that goes into it. In planning the test, consideration must be given to the nature of the objective to be measured, the purpose it is to serve, and the conditions under which it will be used.

1. *Adequate provision should be made for evaluating all the important outcomes of instruction.* A careful statement of the philosophy of the school and the objectives of the particular course should be available from the start. A survey⁷ of a representative sample of 1660 state, county, and city courses of study revealed that only 13 per cent contained no statement of objectives. To be of maximum helpfulness in either teaching or testing, the objectives should be stated as specifically as possible. The expected pupil behavior must be indicated. It is not enough to say that the objective is "good citizenship" or "an integrated personality." These large indefinite terms must be broken down and stated in usable form.

With the list of teaching objectives for the course clearly and specifically stated, the teacher is ready to consider what procedures will be most appropriate for evaluating progress made toward the attainment of each objective. In other words, the teacher attempts to test what he has tried to teach by using techniques best adapted to each objective.

One writer⁸ suggests that the objectives of instruction may be grouped into eight major categories:

- 1 Functional information
- 2 Various aspects of thinking
- 3 Attitudes
- 4 Interests, aims, purposes, appreciations
- 5 Study skills and work habits
- 6 Social adjustment and social sensitivity
- 7 Creativeness
- 8 A functioning social philosophy

Another classification⁹ contains ten major types:

- 1 The development of effective methods of thinking
- 2 The cultivation of useful work habits and study skills
- 3 The inculcation of social attitudes
- 4 The acquisition of a wide range of significant interests
- 5 The development of increased appreciation of music, art, literature, and other aesthetic experiences
- 6 The development of social sensitivity
- 7 The development of better personal social adjustment
- 8 The acquisition of important information

⁷ B. E. Leary, *A Survey of Courses of Study and Other Curriculum Materials Published Since 1934*, Washington: United States Office of Education, 1938.

⁸ Louis F. Rath, "Evaluating the Program of a School," *Educational Research Bulletin* 17, 5784, March 16, 1938.

⁹ Smith, Tyler, and staff, *op cit*, page 18.

- 9 The development of physical health
- 10 The development of a consistent philosophy of life

For any given course these objectives must be expressed in terms of the specific changes in pupils which the teacher is seeking to bring about. A rather detailed inventory of the particular facts, principles, concepts and skills of the course is required, as well as the specific mental processes the pupil is expected to employ. To measure whether these processes are really functioning, the teacher's inventory just mentioned must be presented to the pupils in language different from that of the text and class discussion, and opportunities must be offered to apply or to relate the objectives to new problems and situations. The center of gravity is the behavior of pupils rather than subject matter. The teacher must not confuse ends and means. The true relationship has been stated as follows: "The real ends of instruction are the *lasting* concepts, attitudes, skills, abilities and habits of thought, and the improved judgment or sense of values acquired, the detailed materials of instruction—the specific factual content—are to a large extent only a means toward these ends."¹⁰

A group of English teachers, for example, were able to recognize seven different aspects of "appreciation of literature." They then suggested the following ways in which these aspects of appreciation may manifest themselves in pupil behavior.¹¹

1 *Satisfaction in the Thing Appreciated* Appreciation manifests itself in a feeling on the part of the individual, in keen satisfaction in, and enthusiasm for the thing appreciated. The person who really appreciates a given piece of literature finds in it an immediate persistent, and easily renewable enjoyment of extraordinary intensity.

2 *Desire for More of the Thing Appreciated* Appreciation manifests itself in an active desire on the part of the individual for more of the thing appreciated. The person who really appreciates a given piece of literature is desirous of prolonging, extending, supplementing, renewing his first favorable response toward it.

3 *Desire to Know More about the Thing Appreciated* Appreciation manifests itself in an active desire on the part of the individual to know more about the thing appreciated. The person who really appreciates a given piece of literature is desirous of understanding as fully as possible the significant meanings which it aims to express and of knowing something about the genesis, its history, its locale, its sociological background, its author, etc.

¹⁰ I. F. Inquist, "The Use of Tests in the Accreditation of Military Experience and in the Educational Placement of War Veterans," *Educational Record*, 25: 366, October, 1944.

¹¹ Louis Rath, "Appraising Certain Aspects of Student Achievement," *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I*, pages 114-115. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1938. This reference contains some very suggestive tests designed to measure appreciation of literature, attitudes held toward important social issues, and important aspects of thinking. Also see William A. Brownell (Chairman), "The Measurement of Understanding," *Forty-Fifth Yearbook of the National Society for the Study of Education, Part I* (Chicago: University of Chicago Press, 1946).

- 9 The development of physical health
- 10 The development of a consistent philosophy of life

For any given course these objectives must be expressed in terms of the specific changes in pupils which the teacher is seeking to bring about. A rather detailed inventory of the particular facts, principles, concepts and skills of the course is required, as well as the specific mental processes the pupil is expected to employ. To measure whether these processes are really functioning the teacher's inventory just mentioned must be presented to the pupils in language different from that of the text and class discussion and opportunities must be offered to apply or to relate the objectives to new problems and situations. The center of gravity is the behavior of pupils rather than subject matter. The teacher must not confuse ends and means. The true relationship has been stated as follows: "The real ends of instruction are the *lasting* concepts, attitudes, skills, abilities and habits of thought, and the improved judgment or sense of values acquired, the detailed materials of instruction—the specific factual content—are to a large extent only a means toward these ends."¹⁰

A group of English teachers for example were able to recognize seven different aspects of 'appreciation of literature.' They then suggested the following ways in which these aspects of appreciation may manifest themselves in pupil behavior:¹¹

1 *Satisfaction in the Thing Appreciated* Appreciation manifests itself in a feeling on the part of the individual in keen satisfaction in and enthusiasm for the thing appreciated. The person who really appreciates a given piece of literature finds in it an immediate, persistent and easily renewable enjoyment of extraordinary intensity.

2 *Desire for More of the Thing Appreciated* Appreciation manifests itself in an active desire on the part of the individual for more of the thing appreciated. The person who really appreciates a given piece of literature is desirous of prolonging, extending, supplementing, renewing his first favorable response toward it.

3 *Desire to Know More about the Thing Appreciated* Appreciation manifests itself in an active desire on the part of the individual to know more about the thing appreciated. The person who really appreciates a given piece of literature is desirous of understanding as fully as possible the significant meanings which it aims to express and of knowing something about the genesis, its history, its locale, its sociological background, its author, etc.

¹⁰ E. F. Lindquist, "The Use of Tests in the Accreditation of Military Experience and in the Educational Placement of War Veterans," *Educational Record* 25, 366, October 1944.

¹¹ Louis Rath, "Appraising Certain Aspects of Student Achievement," *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I*, pages 114-115. Quoted by permission of the Society, Bloomington, Illinois: Public School Publishing Company, 1938. This reference contains some very suggestive tests designed to measure appreciation of literature, attitudes held toward important social issues, and important aspects of thinking. Also see William A. Brownell (Chairman), "The Measurement of Understanding," *Forty-Fifth Yearbook of the National Society for the Study of Education, Part I*, Chicago: University of Chicago Press, 1946.

4 *Desire to Express One's Self Creatively* Appreciation manifests itself in an active desire on the part of the individual to go beyond the thing appreciated, to give creative expression to ideas and feelings of his own which the thing appreciated has chiefly engendered. The person who really appreciates a given piece of literature is desirous of doing for himself, either in the same or in a different medium, something of what the author has done in the medium of literature.

5 *Identification of One's Self with the Thing Appreciated* Appreciation manifests itself in the individual's active identification of himself with the thing appreciated. The person who really appreciates a given piece of literature responds to it very much as if he were actually participating in the life situations which it represents.

6 *Desire to Clarify One's Own Thinking with Regard to the Life Problems Raised by the Thing Appreciated* Appreciation manifests itself in a desire on the part of the individual to clarify his own thinking with regard to specific life problems raised by the thing appreciated. The person who really appreciates a given piece of literature is stimulated by it to rethink his own point of view toward certain of the life problems with which it deals and perhaps subsequently to modify his own practical behavior in meeting these problems.

7 *Desire to Evaluate the Thing Appreciated* Appreciation manifests itself in a conscious effort on the part of the individual to evaluate the thing appreciated in terms of such standards of merit as he himself at the moment, tends to subscribe to. The person who really appreciates a given piece of literature is desirous of discovering and describing for himself the particular values which it seems to hold for him.

Arnold¹² has shown that critical thinking can be taught in the elementary school and that various phases of the process can be measured. His study assumed that critical thinking involves the intelligent use of data, which was defined as the "ability to recognize relevance, dependability, bias in source, and adequacy of data in regard to a particular problem, question, or conclusion." The following item has to do with the recognition of bias in data.

Three boys were talking about whether or not a boy Jim was really "out" in a game of baseball that had been played that afternoon. John was on Jim's side. George was on the other side. Bill was not playing but was watching the game. Which of the three boys is most likely to be right? — — — — —
Why? — — — — —

Recognition of the adequacy of data was measured by such test situations as the following:

Some people were talking about a ball team. Someone asked the others to tell why they thought this team was good. Here are the answers. Read them carefully. Put *B* before the one of these three that you think is the best reason for thinking this is a good ball team. Put *P* before the one you think is poorest, and put *F* before the one you think is just fair.

- 1 I think this is a good ball team
- 2 I have seen them play once, and I think they are good players
- 3 I have seen them play several times, and they are good players

¹² Dwight L. Arnold, "Testing Ability to Use Data in the Fifth and Sixth Grades," *Educational Research Bulletin*, 17, 255-259, 278, December 7, 1938.

Read these next two Find the better reason of the two, place a *B* before it
Find the poorer reason, place a *P* before it

_____ 1 A man who studies baseball and writes much about it said they were a good team

_____ 2 A man I talked to on the street the other day said they were good players

Some boys were talking about a boy they called *D* Jim said, "I saw *D* take a pencil from another pupil's desk. This makes me think he is a thief." If this is all that Jim knew about this, do you think Jim is right in thinking that *D* was a thief?

Put a line under your answer YES NO AM NOT SURE

Now tell why you answered as you did

It must be recognized, moreover, that some of the objectives of instruction cannot be measured by paper-and-pencil tests of this kind. At times rating scales, check lists, and other devices for recording observations of the individual at work or play are required. The term "test" in this discussion includes any instrument that affords valid evidence of progress made by pupils toward the attainment of the objectives of instruction.

One of the least tangible objectives of instruction is *creativity*.¹³ Grimes and Bordin¹⁴ have proposed that creative expression in art should result in the development of certain personality traits. These traits would be evaluated by the art teacher through observation during a conversation with his pupils. This process is guided by a check list upon which the teacher's record is entered. Grimes and Bordin suggest that this technique would be more valuable if a group of teachers co-operated in the construction of a check list of their own, rather than adopted wholesale the list given below.

1 *Initiative*—willingness to go into the unknown, to start off on a new track, to attempt something never attempted before, perseverance after recognizing a dead end, willingness to try again.

1 Attempts a medium, technique, or subject never attempted before.

2 Does not accept as final the view of the subject which he happens to have in the place where he begins, but moves around and views subject to be painted from many angles before deciding where to work.

3 Does not take for granted the posed object to be painted, but views the total situation and sees what, for him and his experience, there is in it that is paintable.

4 Assists in posing the model or arranging still life, and the like.

¹³ For stimulating discussions of this general topic, see Joy P. Guilford "Creativity," *American Psychologist* 5: 441-454 September, 1950, and Louis L. Thurstone "Creative Talent," Chapter 2 in *Applications of Psychology* New York: Harper & Brothers, 1952.

¹⁴ James W. Grimes and Edward Bordin, "A Proposed Technique for Certain Evaluations in Art," *Educational Research Bulletin* 18: 1-5 29 January 4 1939.

- 5 Brings material to be painted as still life objects and the like to the studio from outside sources
- 6 Starts to work rather than depending on the teacher for instructions as to how to proceed
- 7 Does not mind making mistakes but pursues the work despite reverses and difficulties
 - a) Scrapes off in painting in oils the thick paint when canvas is gummed up in water color washes out when these steps are necessary for continued work on the painting
 - b) Does not regard initial drawing for a composition as absolute but moves shapes as the developing experience demands adjustments and changes
- 8 Participates in class discussions and contributes ideas and experiences
- 9 Pursues some meaningful activity as sketching if he finishes before the others in the class rather than stalling around
- 10 Does not demand approval or supervision from the teacher or other students at almost every step in his work
- 11 Places his work away from himself, changes viewpoint in order to get a more objective view of his work

11 *Concentration Interest Motivation*—vigor with which an individual attacks a problem and his oneness of purpose which would result in excluding factors nonrelevant to a given problem (perseverance is implied here)

- 1 Not distracted by others coming and going or talking
- 2 Does not come and go himself
- 3 Does not idly converse about matters exterior to the situation
- 4 Does not let the work of others distract him from his own problems
- 5 Does not quickly come to a dead end in his own work
- 6 Does not stall around pretending to be working on a project
- 7 Does work outside of the class that has bearing on class work, as go to gallery, sketch draw, and consult reproductive material
- 8 Works on painting after hours
- 9 Requests information as to work relative to his development
- 10 Contributes to class discussion
- 11 Talks to friends about his work and attempts to explain what he has been accomplishing and learning
- 12 Paints the same subject more than once does not give out quickly as regards subject matter
- 13 Does not continually consult the time

III *Judgment*—weighing the factors in a situation and taking them into account before initiating a new action that is considering the possible results of an action before the initiation of it seeing the social implications of a proposed action Implied in this is knowing when to go off into the unknown and when not to, knowing when to pursue an independent course and when not to

- 1 Does not aid others to the point of interfering with the progress of his own work or that of others

- 2 Attempts to understand the point of view of others rather than thinking of ways to justify himself
- 3 Selects a position to work which does not obstruct another's view of the subject
- 4 Does not talk so loudly that it distracts others who are working
- 5 Takes into consideration the desires and interests of others when arranging a subject to be painted
- 6 Analyzes his working situation in relationship to time, when there is a group-determined time limit
- 7 Knows when to pursue a project further and when to discard it and start another
- 8 Uses materials and cares for them efficiently—cleans palettes washes brushes, and the like
- 9 Works plastically—that is, he allows for working with the forms rather than setting down an architectural plan as a rigid drawing which is filled in with color. Shows evidence of an exploratory and feeling-out attitude rather than a rigid method of working
- 10 Examines critically criticism by others and makes use of it only so far as he feels it significant, that is, he reacts to it in terms of its validity rather than emotionally
- 11 Takes into consideration the needs of others when using group materials
- 12 Avoids vacillation in following out his own painting rather than shifting in style, execution, and attitude, as he sees others in the class going in a direction different from his own

IV *Co-operation*—the willingness to work in a group as a member of it in relationship to the teacher, individual members, and the whole group

- 1 Makes use of criticism, does not react to it as personal insult, or cry, show anger, or leave class
- 2 Is willing to alter his personal objectives to meet the situation

It must be kept in mind that the objectives of the course represent *directions* of progress rather than *destinations* to be arrived at by individual pupils at any particular time. As far as possible, the progress of each individual should be measured in terms of his own interests, needs, and abilities. This is the aim of the modern school. The degree to which it is actually attained in any particular situation is dependent upon the resources available as well as upon the educational philosophy and skill of the teaching staff.

2 *The test should reflect the approximate proportion of emphasis in the course.* To insure a reasonable balance in the test, it is essential to draw up in outline form a sort of "job analysis" or "table of specifications."¹⁵ This will guide the test maker much as the architect's blueprint and specifica-

¹⁵ For a systematic approach to specifying the content of individual items see John C. Flanagan, "The Use of Comprehensive Rationales in Test Development," *Educational and Psychological Measurement* 11: 151-153, Spring 1951.

tions guide the building contractor. It is well to indicate not only the various objectives the teacher has had in mind but also at least roughly the relative amount of emphasis each objective has received in the actual teaching of the course. For example, the same test might not be equally valid for two teachers of a course in general science using the same textbook. This would be the case if one teacher emphasized almost altogether the memorization of isolated facts while the other was much more concerned with the understanding of facts in relation to other facts and in their application to practical problems in the community. The test should attempt to reflect faithfully the teaching emphasis. The amount of time devoted by the teacher to the various topical divisions of the course is a rough indication of what he considers to be their relative importance. The content of the test should show a similar proportion. The time devoted to a topic can at best indicate only the number of items to be included and not the type of the items. The type of items to be used will depend upon the nature of the objective to be measured. A topical outline is only a partial guide to test construction. The table of specifications should also indicate the approximate teaching emphasis from the standpoint of knowledge, skills, attitudes, and other types of objectives that have been sought.

3 *The nature of the test must take into consideration the purpose it is to serve.* Any test is valid to the degree that it serves a specific purpose. If the purpose of the test is to afford a basis for school marks or for classification it will attempt to rank the pupils in order of their total achievement. But if the purpose of the test is diagnosis its value will depend upon its ability to reveal specific weaknesses in the achievement of individual pupils. Diagnostic tests would cover a limited scope but in much greater detail than a test of general achievement and would be arranged so as to reveal the scores on the separate parts. The range of difficulty of the items and the discriminating value of the items individually are relatively less important in diagnostic tests. This is also true of mastery tests administered at the end of a teaching unit to determine when the minimum essentials have been achieved.

4 *The nature of the test must take into consideration the conditions under which it is to be administered.* In planning the test attention must be given to such factors as the time available for testing, the facilities for duplicating the tests, and the cost of the materials, as well as the age and experience of the pupils being tested.

B Preparing the Test

The second step is the actual preparation of the test. It has been found from experience that the following rules or suggestions are helpful.

1 *The preliminary draft of the test should be prepared as early as possible.* Many teachers find it desirable to jot down items to be included in the tests day by day as the teaching progresses. This is reasonable assurance that

no important point in the course will be omitted in the test. If this is not done, the supplementary material of the course which is not included in the textbook and which may be of unusual value is especially likely to be overlooked. This practice also permits the material to "grow cold" and consequently to be more correctly appraised before it is included in the final draft of the test.

2 *The test may include more than one type of item.* A variety of test types is likely to be more interesting to the pupil than a single form. This is especially true of long tests. Moreover, the requirement that the type of test situation should be the one which is most appropriate to the material to be included will sometimes necessitate that three or four forms of objective items be used. These objective items are frequently combined with one or more discussion questions to make up the test.

3 *Most of the items in the final test should be of approximately 50 per cent difficulty* after being "corrected for chance" by the procedure described in footnote 32 on page 160. That is, about half of the group should "know" the answer to each item, while the other half should not. This requirement cannot be met very closely in the typical school situation, however, because item difficulties in the preliminary form of the test will vary considerably. There will usually be too few items to permit discarding those not close to the 50 per cent mark. A suggested rule-of-thumb method for constructing the final (post try-out) form is this. For motivational purposes, let Items 1 and 2 be so easy that almost nobody will miss them. Put aside (do not use) all other items whose correct answer was "known" by less than 16 per cent or more than 84 per cent of the students in the try-out group. Then let Item 3 be the easiest of the remaining items, one "known" by about 84 per cent of the persons tested. Arrange the other items in ascending order of difficulty, with the hardest ones at the end of the test.

The above discussion implies that for maximum discrimination the difficulty of the entire test should be such that, when allowance is made for chance, the average pupil in the group makes about 50 per cent of the possible score. It is clear, then, that a test which is of ideal difficulty for one class may be too easy or too difficult for other classes.

Perhaps one of the worst defects of most teacher-constructed tests is failure to make the items difficult enough, probably in large measure because of the influence of the "70 per cent is passing" tradition. In order to pass all but a few students, the teacher who grades on a percentage-right basis must build tests for which the average score is 80 per cent or more, thereby causing the items that make up the test to be too easy for efficient measurement.

A few exceptions to this principle should be noted. In speed tests for such subjects as arithmetic and typewriting, where the objective is *rate* rather than *power*, all items should be rather easy. Also, in both mastery and diag-

nostic tests the content is determined primarily by the *importance* of the subject matter rather than its *difficulty*. An adequate diagnostic test in the fundamental combinations in addition, for example, might yield many almost perfect scores in a strong class, and scores well below 50 per cent in a weak class.

4 *It is usually desirable to include more items in the preliminary draft of the test than will be needed in the final form.* This will permit "culling out" later on, items that may appear weak or not needed to produce the proper balance in the test. For each subdivision of the test, from 25 to 50 per cent more items should be prepared than are likely to be required.

5 *After some time has elapsed, the test should be subjected to a critical revision.* Then the items should be carefully checked with the table of specifications to see that the test shows the desired emphasis upon the various topics. A careful reading of the test after an interval of time will usually reveal some objectionable items. It is a good plan to have the test criticized by other teachers of the same subject. In this way some items are likely to be found which cover points of doubtful importance, others which are not clearly stated, and perhaps others about which there is disagreement as to the answers. The wording of the items should receive critical attention, particularly to avoid ambiguity. One serious error is the wording of items so that more than one reasonable interpretation is possible. The trouble with such ambiguous items is that a certain answer is correct with one interpretation, but with another interpretation a different answer is reasonably correct.

6 *The items should be so phrased that the content, rather than the form of the statement, will determine the answer.* A common mistake is to include a telltale word or phrase that affords an unwarranted clue to the answer. These so-called *specific determiners* are especially common in true-false items. It has been found that statements containing emphatic words, such as the adverbs "always," "never," "entirely," "absolutely," "exclusively," and the like, are much more likely to be false than true. On the other hand words or expressions that limit the statement, such as "may," "sometimes," "as a rule," "in general," and the like, are much more likely to be true than false. Either these expressions should be avoided entirely, a suggestion which is rarely feasible, or items containing them should be carefully balanced so that approximately the same number are true as false. Avoiding the language of the text will prevent pupils with good rote memories from answering items they may not understand. Sometimes clues are afforded by the spelling or by the grammatical form of the item. It is not unlikely that one of the reasons why many pupils prefer objective tests to other types is that such tests often contain items so worded as to be answered from a minimum knowledge of the subject matter involved. Such defects, however, are not inherent in objective testing, they can be avoided by the

alert test maker Administering the test to persons unfamiliar with the content of the course will often reveal those items which can be answered from general intelligence or from a general knowledge of language forms and usage

The opposite mistake is often made also Figurative language, needlessly heavy vocabulary, or involved sentence structure may so obscure the meaning of an item that it is marked incorrectly by pupils who really understand the point Bob Burns' story of the time Grandpa Snazzy was a witness in court illustrates this error

The attorney says "Now, Mr Snazzy, did you or did you not, on the date in question or at any time previously or subsequently, say or even intimate to the defendant or anyone else, whether friend or mere acquaintance or in fact a total stranger, that the statement imputed to you, whether just or unjust and denied by the plaintiff, was a matter of no moment or otherwise? Answer—did you or did you not?"

Grandpa thinks a while and then says, "Did I or did I not what? "

Unless the test aims specifically to measure reading ability or general intelligence, the form of the item should neither impose unreasonable obstacles in the pupil's way nor provide clues which are too obvious Both defeat the purpose for which the test was intended

7 *The items should be so worded that the whole content functions in determining the answer, rather than only a part of it* There is often a wide discrepancy between what actually determines the pupil's response to a test and what the teacher intended One of the principal reasons for this discrepancy is that only a part of the content of the item functions, the rest being wholly inert as far as the pupil is concerned Lindquist¹⁶ gives some excellent examples of this difficulty Two of these, shown below, should make the problem clear Note the first

The leader in the making of the compromise tariff of 1833 was (1) Clay, (2) Webster, (3) Jackson (4) Taylor, (5) Harrison

That the majority of the pupils who responded to this item correctly did so on the superficial basis of the strong verbal association between the words "compromise" and "Clay" is evidenced by the fact that fewer than half of them responded correctly when the item appeared in the following form

The leader in the tariff revision of 1833 was (1) Clay, (2) Webster, (3) Jackson, (4) Taylor, (5) Harrison

That the matching type of test is also subject to this error is shown by the next illustration

¹⁶ Herbert E Hawkes E F Lindquist and C R Mann *The Construction and Use of Achievement Examinations* pages 73-81 Boston Houghton Mifflin Company, 1936

Directions Below are two columns of items. Match the items in the two columns by placing on the line before each group of words in Column A the right number from Column B.

Column A

- 1 a Phoenician contribution to civilization
- 2 most famous building of the ancient Greek world
- 3 the fleet whose defeat in 1588 gave England the control of the Atlantic Ocean
- 4 a boundary between two colonies that later became famous as the division between free and slave territory
- 5 the victory which caused France to come to our aid during the Revolutionary War
- 6 the law that forbade slavery north of the Ohio River
- 7 a ruling by the Supreme Court which opened all territory to slavery

Column B

- 1 Mason and Dixon Line
- 2 Spanish Armada
- 3 Saratoga
- 4 Dred Scott Decision
- 5 Parthenon
- 6 Missouri Compromise
- 7 Alphabet
- 8 Printing Press
- 9 Ordinance of 1787

In most of the above items a single word gives the clue. For example "boundary" in item 4 suggests "line" in response 1. Likewise either "ruling" or "court" in item 7 suggests "decision" in response 4. If a pupil knows that "armada" means "fleet," he would be able to match item 3 with response 2 without knowing the date, the country or the event involved. It should be noted, furthermore, that probably the above test would still be poor, even if each item were well worded, because the items included are so diverse in character.

The test maker should attempt to anticipate the specific mental processes the pupil will employ in each response. For each item the teacher should raise such questions as the following: Are there any parts of the item that the pupil may disregard entirely and yet respond correctly? What is the minimum amount of knowledge required for a correct response?

8 All the items of a particular type should be placed together in the test. Sometimes completion, true-false, and multiple-choice items of varying numbers of choices are thrown together in random order. This arrangement is rarely, if ever, desirable. It is good practice to place together the items of similar type. Such an arrangement not only facilitates the scoring of the test and the interpretation of the scores, but enables the pupil to take full advantage of the mind set imposed by a particular item form.

9 The items in the test should be arranged in ascending order of difficulty. It is especially important to have the easiest items at the beginning and the hardest ones at the end of the test. It will be recalled that one of the problems of measurement is to arrange conditions so that the thing being

measured is disturbed as little as possible in the act of measuring. The psychological justification for placing the easiest items first is that such an arrangement has a wholesome effect upon the morale of the pupils taking the test. On the other hand, placing very difficult items at the beginning is likely to produce needless discouragement in the pupils, particularly with those of average ability and below. If the most difficult items come toward the end of the test, only the more capable pupils will probably get to them. After all, the only function of such items is to discriminate among the high-ranking pupils. In any event, any disturbing influence on the weaker pupils will come too late to affect the results seriously.

In advance of an actual tryout of the test, it is impossible to determine anything more than a rough estimate of the true difficulty order of the items, unless one is willing to go to the trouble of obtaining the pooled judgment of three or more persons.¹⁷ The judgment of a single experienced teacher regarding the difficulty of the items is likely to have some validity. In any case it is usually possible to pick out those items that will be at the extremes of the scale, and fortunately this is what is needed most. In later revisions of the test, the items can be placed in more exact order of difficulty.

10 *A regular sequence in the pattern of responses should be avoided.* The order of correct responses should be a chance order rather than a regular pattern.¹⁸ If items are arranged alternately true and false, or two true and two false, for example, the pupil is likely to discover the arrangement. To facilitate scoring, it is sometimes suggested that multiple-choice items be so arranged that the option numbers of the correct responses give combinations easy to remember, such as a date like 1453. But there is always risk that the pupil will "get the hang" of the pattern and answer successfully without considering the content of the item at all.

11 *Provision should be made for a convenient written record of the pupil's responses.* Such a record is a check list, a rating scale, or some other similar form upon which the observer makes a systematic and permanent record of a pupil's behavior under a given set of conditions. It is particularly difficult to provide a satisfactory written record of responses on oral quizzes.¹⁹ In the ordinary test the pupil makes his own record in writing either on the test paper or a specially prepared answer sheet. The problem then is merely that of arranging the test so that the labor of scoring will be reduced to a minimum. Such devices as numbering or lettering the re-

¹⁷ Sherman Tinkelman, *Difficulty Prediction of Test Items*, page 49. New York: Bureau of Publications, Teachers College, Columbia University, 1947. Further investigation of this problem was undertaken by Irving Lorge and Lorraine Kruglov, "A Suggested Technique for the Improvement of Difficulty Prediction of Test Items," *Educational and Psychological Measurement*, 12, 554-561, Winter 1952.

¹⁸ Scarvia B. Anderson, "Sequence in Multiple Choice Item Options," *Journal of Educational Psychology*, 43, 364-368, October 1952.

¹⁹ Max M. Kestick and Belle M. Nixon, "How to Improve Oral Questioning," *Peabody Journal of Education*, 30, 209-217, January 1953.

responses in multiple-choice items and the blanks in completion items so that the responses will be recorded in a column rather than scattered irregularly over the page, save time and reduce the chances of error in scoring. Merely grouping the items by fives rather than spacing them uniformly, reduces somewhat the eyestrain in scoring the test.

12 *The directions to the pupil should be as clear, complete, and concise as possible.* The aim should be to make the instructions so clear that the weakest pupil in the group knows what he is expected to do although he may not be able to do it. The pupil should be told how and where to mark the items, the time allowed to do so and any reduction for errors to be made in scoring. The amount of detail required will depend upon the maturity of the pupils and their experience with that particular type of test. To very young children, for example, it will be better to say "draw a line under" rather than "underline" and "draw a ring around the right answer" rather than "encircle the correct response." In the lower grades it is usually desirable for the teacher to read the directions aloud to the pupils while they follow silently the written directions on their test papers. Wherever the form of the test is unfamiliar or complicated a generous use of samples correctly marked and fore-exercises or practice tests that do not count in determining the score is to be recommended. Sometimes a blackboard demonstration is the best way to make the procedure clear. As the pupils become familiar with the various types of items and the procedure used in scoring them, the directions may be greatly abridged.

A single illustration should make these points clear. The following directions may be considered reasonably satisfactory for a class unfamiliar with objective tests.

DIRECTIONS TO THE PUPIL Below are thirty statements about measurement in education. Examine each statement and decide whether it is true or false. In the () before each statement you think is true put + in the () before each statement you think is false put 0. You will have ten minutes for the test. Your score will be the number right minus the number wrong. You may mark an answer even when you have only a slight hunch, but do not guess wildly. Study the samples below. They are answered correctly.

SAMPLES

- (0) A. High reliability insures high validity in a test.
(+) B. Group tests of intelligence originated in America.

After the pupils have become familiar with true-false tests and the method employed in scoring them, the directions may be shortened to a form somewhat as follows.

DIRECTIONS In the () before each item put + if true and 0 if false. You will have ten minutes for the test.

One other point warrants consideration. Should pupils be told or encouraged to guess at items about whose answers they are in doubt? Some authorities would require the pupils to attempt all items on recognition

measured is disturbed as little as possible in the act of measuring. The psychological justification for placing the easiest items first is that such an arrangement has a wholesome effect upon the morale of the pupils taking the test. On the other hand, placing very difficult items at the beginning is likely to produce needless discouragement in the pupils, particularly with those of average ability and below. If the most difficult items come toward the end of the test, only the more capable pupils will probably get to them. After all, the only function of such items is to discriminate among the high-ranking pupils. In any event, any disturbing influence on the weaker pupils will come too late to affect the results seriously.

In advance of an actual try out of the test, it is impossible to determine anything more than a rough estimate of the true difficulty order of the items, unless one is willing to go to the trouble of obtaining the pooled judgment of three or more persons.¹⁷ The judgment of a single experienced teacher regarding the difficulty of the items is likely to have some validity. In any case it is usually possible to pick out those items that will be at the extremes of the scale, and fortunately this is what is needed most. In later revisions of the test, the items can be placed in more exact order of difficulty.

10 *A regular sequence in the pattern of responses should be avoided.* The order of correct responses should be a chance order rather than a regular pattern.¹⁸ If items are arranged alternately true and false, or two true and two false, for example, the pupil is likely to discover the arrangement. To facilitate scoring it is sometimes suggested that multiple-choice items be so arranged that the option numbers of the correct responses give combinations easy to remember, such as a date like 1453. But there is always risk that the pupil will "get the hang" of the pattern and answer successfully without considering the content of the item at all.

11 *Provision should be made for a convenient written record of the pupil's responses.* Such a record is a check list, a rating scale, or some other similar form upon which the observer makes a systematic and permanent record of a pupil's behavior under a given set of conditions. It is particularly difficult to provide a satisfactory written record of responses on oral quizzes.¹⁹ In the ordinary test the pupil makes his own record in writing either on the test paper or a specially prepared answer sheet. The problem then is merely that of arranging the test so that the labor of scoring will be reduced to a minimum. Such devices as numbering or lettering the re-

¹⁷ Sherman Tinkelman, *Difficulty Prediction of Test Items*, page 49. New York: Bureau of Publications, Teachers College, Columbia University, 1947. Further investigation of this problem was undertaken by Irving Lorge and Lorraine Kruglov, "A Suggested Technique for the Improvement of Difficulty Prediction of Test Items," *Educational and Psychological Measurement* 12: 554-561, Winter 1952.

¹⁸ Scarvia B. Anderson, "Sequence in Multiple-Choice Item Options," *Journal of Educational Psychology* 43: 364-368, October 1952.

¹⁹ Max M. Kostick and Belle M. Nixon, "How to Improve Oral Questioning," *Peabody Journal of Education* 30: 209-217, January 1954.

sponses in multiple-choice items and the blanks in completion items, so that the responses will be recorded in a column rather than scattered irregularly over the page, save time and reduce the chances of error in scoring. Merely grouping the items by fives, rather than spacing them uniformly, reduces somewhat the eyestrain in scoring the test.

12 *The directions to the pupil should be as clear, complete, and concise as possible.* The aim should be to make the instructions so clear that the weakest pupil in the group knows what he is expected to do although he may not be able to do it. The pupil should be told how and where to mark the items, the time allowed to do so, and any reduction for errors to be made in scoring. The amount of detail required will depend upon the maturity of the pupils and their experience with that particular type of test. To very young children, for example, it will be better to say "draw a line under" rather than "underline," and "draw a ring around the right answer" rather than "encircle the correct response." In the lower grades it is usually desirable for the teacher to read the directions aloud to the pupils while they follow silently the written directions on their test papers. Wherever the form of the test is unfamiliar or complicated, a generous use of samples correctly marked and fore-exercises or practice tests that do not count in determining the score is to be recommended. Sometimes a blackboard demonstration is the best way to make the procedure clear. As the pupils become familiar with the various types of items and the procedure used in scoring them, the directions may be greatly abridged.

A single illustration should make these points clear. The following directions may be considered reasonably satisfactory for a class unfamiliar with objective tests.

DIRECTIONS TO THE PUPIL Below are thirty statements about measurement in education. Examine each statement and decide whether it is true or false. In the () before each statement you think is true, put +, in the () before each statement you think is false, put 0. You will have ten minutes for the test. Your score will be the number right minus the number wrong. You may mark an answer even when you have only a slight hunch, but do not guess *wildly*. Study the samples below. They are answered correctly.

SAMPLES

- (0) A High reliability insures high validity in a test.
(+) B Group tests of intelligence originated in America.

After the pupils have become familiar with true-false tests and the method employed in scoring them, the directions may be shortened to a form somewhat as follows:

DIRECTIONS In the () before each item put + if true, and 0 if false. You will have ten minutes for the test.

One other point warrants consideration. Should pupils be told or encouraged to guess at items about whose answers they are in doubt? Some authorities would require the pupils to attempt all items on recognition

tests They would include some such statement as, 'If you do not know guess!' Others would go to the other extreme and say, 'Do not guess!' Still others perhaps the majority, would be content with informing the pupil that the correction formula²⁰ is to be employed and let him use his judgment about attempting doubtful items Unfortunately, the experimental evidence on this point is neither extensive nor altogether convincing Most of the studies have merely attempted to compare the relative effect of the first two practices upon the validity and reliability of the scores without considering the third possibility at all

The results have usually favored do-not-guess *wildly* instructions by a slight margin Davis is particularly opposed to forced guessing²¹

To force students to mark answers to items based on reading passages available for reference is undesirable but to force them to mark answers to items testing specific subject matter that they do not know and cannot figure out is far worse It is not only frustrating to the students but it goes contrary to good teaching practices and compels the students to break habits of carefulness that the schools try hard to inculcate Once in a while it is possible that students might be told that as part of an experiment they are required to mark every item in a test even when they have no idea what to mark but in systematic testing programs this would be inadvisable as well as impractical It might eliminate variations in the number of omissions and thus wipe out some of the effects of differences in personality but it would do this at a cost of antagonizing teachers and frustrating students It would also introduce additional chance variance into the scores

However Votaw²² Lentz²³ Cronbach²⁴ and others have found some evidence of a factor of 'acquiescence' operating in taking tests since do-not guess instructions placed ascendant students at an advantage over submissive students It is argued that such instructions reduce the validity of achievement tests since they become in some degree measures of personality traits Some investigators have also found that good students tend to improve their scores when they attempt doubtful items, whereas poor students do not

All things considered the authors offer the following recommendations

- a The use of multiple-choice tests with fewer than four responses to each item should be avoided wherever possible

²⁰ This formula is discussed on pages 156-158

²¹ Frederick B Davis Item Selection Techniques in E F Lindquist (Editor) *Educational Measurement* pages 274-275 Washington D C American Council on Education 1931

²² David F Votaw The Effect of Do-not-guess Directions upon the Validity of True-false or Multiple-choice Tests *Journal of Educational Psychology* 27 698-703 December 1936

²³ Theodore F Lentz Acquiescence as a Factor in the Measurement of Personality *Psychological Bulletin* 30 609 November 1938

²⁴ Lee J Cronbach Studies of Acquiescence as a Factor in the True-false Test *Journal of Educational Psychology* 33 401-413 September 1942 Further Evidence on Response Sets and Test Design *Educational and Psychological Measurement* 10 3-31 Spring 1950

b. Regardless of the number of possible responses, the score should probably be the number right on all tests used with pupils below the junior-high-school level."

c. When tests with only two or three possible responses to each item are used with pupils above the sixth grade, the correction formula should be employed.

d. Whenever the correction formula is to be used, the pupils should be so informed.

e. The reason for the correction formula should be discussed with pupils before the test begins. They should then be allowed to use their best judgment, without being specifically advised by the directions to guess or not to guess.

C. Trying Out the Test

After the test has been prepared according to plan, it is ready to be given a trial in actual use. Since it is impossible in advance to know exactly how good the test is or to locate all the poor items, the tryout should be considered a necessary step in constructing the test in its final form. The following four principles should govern the tryout. With the possible exception of the second, these principles are equally applicable to the later use of the test in its final form.

1. *Every reasonable precaution should be taken to insure normal conditions for the test.* This is important because the responses to any test are partly determined by the conditions under which it is given as well as by the test itself. It is usually well to have the test administered to the pupils in the familiar environment of their own classroom. Any tendency to cheat should be forestalled by careful supervision. Where cheating is likely to be a special problem, pupils may be so seated that every other seat is vacant, or the test items may be arranged in different orders for pupils seated close together.

2. *The time allowance for the test should be generous.* This is more important in the tryout than in the later use of the test in its final form. One reason for this is that the items are arranged at best in only a rough order of difficulty, and, if the time allowance is too short, pupils may not have time to try items toward the end of the test, which they may be capable of answering correctly. Short time allowances should be avoided, therefore in order to secure the data needed for determining the difficulty and the discriminating value of the items. What time allowance is to be considered generous will depend upon the purpose of the test and upon the ability and experience of the pupils. For example, it is obvious that the time limits of speed tests should be so short that even the best pupil does not have time to finish the test. On the other hand, more time should be allowed for diag-

* Admittedly this decision is based on practical considerations rather than on experimental evidence. It is usually difficult to explain to young children the logic of the formula and they are likely to be suspicious of what they do not understand.

nostic tests than for tests of general achievement, and tests of a purely factual character can be answered more quickly than those involving the higher mental processes

Landquist suggests that in general achievement tests, the time allowance should be so adjusted that "at least 75 per cent of the pupils will have time at least to *consider* all items in each section"²⁶ Ruch seemed to favor time limits "so that 90 per cent can attempt all items within their power"²⁷ In accordance with this standard, Ruch suggested that for fairly short items of a factual character, three recall or four recognition items per minute is a "reasonable expectancy for upper-elementary and high-school pupils" For reasoning tests the corresponding time allotments would be increased for recall items to one or two items per minute, and for multiple-choice items to two or three items per minute Younger pupils and longer or harder items would demand still more time

The above standards have in mind the requirements for the ordinary use of the test in its final form, rather than for the tryout, for which more time should be allowed Since so many factors influence the time demands of a particular test, the writers suggest that in the tryout sufficient time be allowed so that all, or almost all, the pupils have time to finish If the examiner will record during the progress of the test the percentage of the pupils who are still at work after various amounts of time have elapsed, this information will be useful in determining the time allowances for later revisions of the test

3 *The scoring procedure adopted should be fairly simple* As a rule, the best procedure in scoring objective tests is to give one point of credit for each correct response In multiple-choice tests this means one point for each item properly marked, and in recall tests it means one point for each blank correctly filled It is unnecessary to weight the items according to estimated difficulty or importance Even in essay examinations weighting is much less important than is ordinarily assumed Almost all pupils will be in the same rank order regardless of the weighting of the individual items²⁸

The correction-for-chance formula actually corrects for omissions rather than for guessing alone If no items are omitted by the students, or if they all omit the same number of items, their relative scores will be the same regardless of whether or not it is employed The general formula is usually written

$$S = R - \frac{W}{O - 1}$$

²⁶ Herbert E. Hawkes, F. F. Landquist and C. R. Mann, *op cit*, page 116

²⁷ G. M. Ruch, *The Objective or New-Type Examination*, page 312 Chicago: Scott Foresman & Company, 1929

²⁸ Cf. Alexander J. Phillips, "Further Evidence Regarding Weighted Versus Unweighted Scoring of Examinations," *Educational and Psychological Measurement*, 3: 151-155, Summer, 1943

In this formula

S is the score corrected for guessing

R is the number of right responses

W is the number of wrong responses not counting omitted items

O is the number of options presented for each item

For two-option or true-false items this becomes

$$S = R - W$$

For three-option items the formula is

$$S = R - \frac{1}{2}W$$

For four-option items the formula is

$$S = R - \frac{1}{3}W$$

For five-option items it is

$$S = R - \frac{1}{4}W$$

If the items have six or more options, it is probably not worth while to correct the "rights" scores for chance

The above formulas are designed to reduce to zero the score of each person who is totally ignorant concerning the material as presented in the test and who guesses at the right answers with a degree of success dependent upon the number of options each item has. For example, if the test contains 100 true-false items and if the testee guesses at each of these he should on the average answer about 50 items "correctly" since there is one chance in two that he will mark any item right purely by accident. Thus the expected score for a wholly uninformed person would be 50 right and 50 wrong. However, he richly deserves a final score of zero which represents his knowledge of the material covered so from the 50 rights are subtracted the 50 wrongs $R - W = 50 - 50 = 0$.

On the other hand, if he answers 50 items correctly and omits the other 50, his score will be $50 - 0 = 50$. Presumably, had he tried the 50 omitted items he would have answered half of them (25) correctly by chance and missed the other 25, making his rights score 50 known + 25 guessed = 75 and his wrongs score 25 $75 - 25 = 50$ the same score he would have secured without any guessing. There is a fallacy in this argument, though, for a student is not likely to know the answers to half the questions and be absolutely ignorant concerning the other half. More likely, he has various degrees of partial information and misinformation concerning a considerable number of items. In many test situations these two types of information seem to cancel each other out, thereby making the $R - \frac{W}{O-1}$ formula suitable.²⁹

²⁹ Frederick B. Davis *op cit* page 271

It should be emphasized that, strictly speaking, the correction formula is needed only when some students have omitted a fairly large number of items, while others have omitted few. Otherwise, the ranks of the students will be unchanged, regardless of whether or not their scores are corrected for "chance." For psychological reasons the teacher may report corrected scores to the students, even though few items have been omitted. This is especially advisable with true-false tests, where the poorest students may not realize the extent of their ignorance and hence may protest if given low grades unless the $R - W$ formula is employed.

If all individuals tested answer every item, the standard deviation of scores corrected for chance is $\frac{O}{O-1}$ times the standard deviation of the rights scores. From this relationship it is apparent that correcting total scores from a "do-guess" true-false test doubles their standard deviation. $O - (O - 1) = 2$ divided by $(2 - 1) = 2$. Likewise, if there are no omissions the standard deviation of corrected-for-chance five-option-item test scores is $\frac{5}{4} = 1\frac{1}{4}$ times the standard deviation of the uncorrected scores.

The authors recommend using the correction formula with true-false and two- and three-option multiple-choice test scores, even when omissions are negligible, in order to emphasize the range of knowledge within the group tested.

4 *Before the actual scoring begins, answer keys and scoring rules should be prepared.* In teacher made objective tests satisfactory scoring keys can be prepared by simply filling in the correct responses, preferably with a colored pencil, on one of the unused tests. Scoring then consists of comparing the pupil's responses with those on the key placed beside his paper. In essay examinations the key consists of a model paper containing a complete set of answers, together with the points to be allowed on each. Definite rules are necessary to secure uniformity in scoring. The rules for scoring objective tests usually say merely that one point will be allowed for each correct response and that no fractional credits will be allowed, and indicate whether or not the correction formula will be used. The rules for essay examinations give the weight for each question, and tell whether or not any deductions are made for errors in spelling, language usage, and so forth. In mathematics tests the rules should cover such points as whether or not the answers must be reduced to lowest terms, whether or not credit will be allowed for solutions correct in principle but with the wrong answer, and the like.

If the students' answers have been recorded on a special answer sheet, such as the International Business Machines general purpose answer sheet, then a punched-out cardboard key will speed up scoring a great deal. One of the most useful of the various answer sheets is shown in Figure 5.²²

²² Answer sheets and the key punch may be purchased from International Business Machines Corporation, Endicott, New York.

D. Evaluating the Test

After the papers have been scored the results should be interpreted and evaluated from two points of view first, as to the quality of the test itself, and, second, as to the quality of the pupils' responses. While the ultimate interest of the test maker is in the light thrown by the test results upon the quality of the teaching and organization that exists in the school, his first

NAME	SCHOOL	DATE	SCORES	DATE OF BIRTH		GRADE OR CLASS	INSTRUCTOR	NAME OF TEST	PAGE			
				1	2							
1									31	61	91	121
2									32	62	92	122
3									33	63	93	123
4									34	64	94	124
5									35	65	95	125
6									36	66	96	126
7									37	67	97	127
8									38	68	98	128
9									39	69	99	129
10									40	70	100	130
11									41	71	101	131
12									42	72	102	132
13									43	73	103	133
14									44	74	104	134
15									45	75	105	135
16									46	76	106	136
17									47	77	107	137
18									48	78	108	138
19									49	79	109	139
20									50	80	110	140
21									51	81	111	141
22									52	82	112	142
23									53	83	113	143
24									54	84	114	144
25									55	85	115	145
26									56	86	116	146
27									57	87	117	147
28									58	88	118	148
29									59	89	119	149
30									60	90	120	150

Figure 5 IBM General Purpose Answer Sheet

concern should be the quality of the test used. Only tests of high merit afford suitable information regarding the school situation. To what extent, then, does the test possess the three characteristics of satisfactory measuring instruments, validity, reliability, and usability? Only the last of these can be confidently determined in advance. The five principles that follow are suggested for evaluating the test from the viewpoint of its validity and reliability.³¹ If the test is found to possess these qualities in high degree, the scores should then be carefully analyzed for their value in instruction and school administration. If the test is found to lack these qualities, the scores can be disregarded and the test subjected to a thorough revision. No matter how carefully the test is prepared in the first place, its merits should be established and not merely assumed.

1 *The difficulty of the test is a rough indication of its adequacy.* The difficulty of the test as a whole is determined by finding what percentage the average score made (corrected for chance, if appropriate) is of the maximum possible score. In general achievement tests, the nearer this average is to 50 per cent, the better. The difficulty of the individual test items is obtained by finding the percentage of successful responses for each item, usually corrected for "chance" in the same manner as the total score.³² Items answered by 100 per cent or by 0 per cent of the pupils are of no value in a test of general achievement. The difficulty of the test is relatively unimportant in mastery tests and in diagnostic tests.

2 *The internal consistency of the individual items in the test is determined by their ability to discriminate between pupils who rank high and those who rank low on the test as a whole.* There are several methods of determining this. Only the simplest of these methods are practical for use with informal tests. A satisfactory procedure for the classroom teacher is to determine the number of correct responses (or of incorrect responses) to each test item by the pupils who rank in the highest 27 per cent of the class on the test as a whole and to compare this with the corresponding number in the

³¹ For a brief but suggestive report see Ellis Weitzman and Walter J. McNamara, "Techniques Used in Analyzing the Learning Achievement of Naval Aviation Cadets," *Journal of Educational Psychology*, 35, 181-185, March 1944.

³² Proportion of testees who know the answer to an item

$$= \left(\text{number who marked it correctly} - \frac{\text{number who marked it incorrectly}}{\text{number of options item has minus one}} \right)$$

divided by the total number of testees. In symbols

$$p = \frac{R - \frac{W}{O-1}}{N}$$

This formula is appropriate only when nearly all examinees have had time enough to try the item. Otherwise see Frederick B. Davis, "Item Selection Techniques," pp. 278-280 in F. F. Lindquist (Editor), *Educational Measurement*, Washington, D. C., American Council on Education, 1935.

lowest 27 per cent of the class.³³ The items in which the number of correct responses of the high group exceeds that of the low group by the largest amount are best, those in which numbers are the same are useless and those in which the number of correct responses of the high group falls behind that of the low group are detrimental. Items showing zero or negative discrimination should be either reworded or thrown out altogether.³⁴

3 *It is a good practice to have the items interpreted or criticized by persons who have taken the test.* It is impossible to anticipate fully all the mental processes pupils will employ in responding to a test item. These can be determined only by making inquiry of pupils who have taken the test. In this way irrelevancies and ambiguities will be revealed that were wholly unsuspected by the maker of the test. Often a slight change in wording is sufficient to remedy the difficulty. At other times the item must be entirely discarded. If a test contains too many of these items the scores on the test should not be counted in determining the pupil's record in the class. Inviting members of the class to assist in this critical evaluation of the test may help to create a favorable attitude toward the measurement process employed by the instructor and is a valuable educational experience in itself.

4 *Whenever possible the results on the test should be checked against an outside criterion.* For short tests covering small units of subject matter this process is likely to be difficult and of little value. Even here it is sometimes helpful to compare the ranks of the pupils on the test with those assigned by the teacher before the test is given. The validity of the longer and more important tests can be determined in a more satisfactory manner by comparing the scores of the pupils on each test with their scores on a good standard test covering the same material and given at about the same time. The coefficient of correlation obtained between the two series of scores is the most exact method of expressing the amount of agreement although a rough indication can be obtained by comparing the percentage of scores which lie in the same fourths of the two series of scores.

5 *It is sometimes desirable to obtain the reliability coefficient of the test.* The authors recognize that it is easy to overestimate the value of the reliability coefficient. The makers of standardized tests have often made this mistake. However the reliability coefficient does have some merit in evaluating informal tests although the value is mainly negative. Low reliability coefficients indicate tests of doubtful merit but high reliability coefficients per se do not establish the value of the tests. To be of real value these coefficients must be supported by other criteria.³⁵

³³ The reasons for selecting contrasting groups from the 27 per cent at the extremes of the distribution are given by Truman L. Kelley. The Selection of Upper and Lower Groups for the Validation of Test Items. *Journal of Educational Psychology* 30: 17-24. January 1939.

³⁴ A complete item analysis of a classroom test is set forth in Appendix B, pages 436-453.

³⁵ In Appendix B, page 452, a simplified method for securing a one-form reliability coefficient is applied to a typical classroom examination.

The construction of an informal teacher-made test, then, involves these four steps: planning, preparing, trying out, and evaluating. It is perhaps more correct to say that these activities constitute a cycle in the construction of a test, for it is often necessary to repeat these steps, particularly the last three, several times before the test is brought to its finished form.

SELECTED REFERENCES FOR FURTHER READING

- Adkins, Dorothy C., and others, *Construction and Analysis of Achievement Tests* Washington, D C U S Government Printing Office, 1947 292 pages
- Cronbach, Lee J., *Essentials of Psychological Testing* New York Harper & Brothers, 1949 475 pages
- Davis, Frederick B., "The AAF Qualifying Examination," *Army Air Forces Aviation Psychology Research Report No 6* Washington, D C U S Government Printing Office, 1947 266 pages
- Gardner, Eric F., "Development and Applications of Tests of Educational Achievement in Schools and Colleges," *Review of Educational Research*, 23 85 101, February, 1953
- Goheen, Howard W., and Kavruck, Samuel *Selected References on Test Construction Mental Test Theory, and Statistics, 1929-1949* Washington, D C U S Government Printing Office, 1950 209 pages
- Goodenough, Florence L., *Mental Testing Its History, Principles, and Applications* New York Rinehart & Company, 1949 Chapter 8, 'The Analysis and Selection of Test Items'
- Jordan, A M., *Measurement in Education An Introduction* New York McGraw-Hill Book Company, 1953 Chapter 2, "Characteristics of Measuring Instruments"
- Lindquist, E F., "Preliminary Considerations in Objective Test Construction' Chapter 5 in E F Lindquist (Editor), *Educational Measurement* Washington D C American Council on Education, 1951
- Odell, C W., *How to Improve Classroom Testing* Dubuque, Iowa William C Brown Company, 1953 Chapter IV, "Test Construction General"
- Stanley, Julian C. 'A Simplified Item-Analysis Procedure,' *American Psychologist*, 6 369 July 1951
- Stanley, Julian C., "A Simplified Method for Estimating the Split-Half Reliability Coefficient of a Test," *Harvard Educational Review*, 21 221-224, Fall, 1951
- Stanley, Julian C., "'Psychological' Correction for Chance," *Journal of Experimental Education*, 22 297-298 March, 1954
- Travers, Robert M W., *How to Make Achievement Tests* New York Odyssey Press, 1950 180 pages
- Travers, Robert M W., "Rational Hypotheses in the Construction of Tests," *Educational and Psychological Measurement*, 11 128-137, Spring, 1951
- Traxler, Arthur E., Jacobs Robert, Selover, Margaret, and Townsend, Agatha, *Introduction to Testing and the Use of Test Results in Public Schools* New York Harper & Brothers, 1953 Chapter 2, "What Do Tests Contribute to Understanding the Individual Pupil?"
- Weitzman, Ellis and McNamara, Walter J., *Constructing Classroom Examinations — A Guide for Teachers* Chicago Science Research Associates, 1949 Chapters 1 and 2 "Basic Aspects of Achievement Tests" and 'Steps in Classroom Testing'

6

Principles of Constructing Specific Types of Objective Tests

A. Introduction

Types of objective tests. The principal types of objective test items used by classroom teachers may be listed as follows

- 1 Recall types
 - a Simple-recall
 - b Completion
- 2 Recognition types
 - a *More common*
 - (1) Alternative-response
 - (2) Multiple-choice
 - (3) Matching
 - b *Less common*
 - (1) Rearrangement
 - (2) Identification
 - (3) Analogy
 - (4) Incorrect statement

This chapter will consider the uses and limitations of the commonly used forms of objective tests and suggest rules which have been found to be of value in constructing them. It will also give illustrative items in a variety of fields, drawn mainly from standard tests.

Frequency of use by teachers. Two early studies present data on the frequency of use by classroom teachers of various forms of test items. In the first of these studies Conneau¹ analyzed 45,418 test items that appeared

¹ Summarized by G. M. Ruch, *The Objective or New-Type Examination*, pages 188-190. Chicago: Scott, Foresman & Company, 1929.

in 375 objective examinations submitted in a prize contest. This study doubtless represented the practice of superior teachers in 1928, rather than that of average teachers. In 1936 Lee and Segel² reported an analysis of the types of informal tests used by 1,600 high school teachers distributed widely over the United States. That there is rather surprising agreement between these studies is indicated by Table 25. In both studies the comple-

TABLE 25

RANKINGS OF TEST ITEMS ACCORDING TO FREQUENCY OF USE AS REVEALED BY TWO STUDIES

<i>Type of Item</i>	<i>Conneau</i>	<i>Lee and Segel</i>
Completion	1	1
True-false	2	2
Multiple-choice	3	4
Essay	11	5
Problem	7	6
Matching	4	7

tion form ranks first and the true-false second. Conneau grouped all recall forms under completion, while Lee and Segel separated out the one-word answer type. This type of item, which ranked third in the latter study, has not been included in the table. The next most popular item is the multiple-choice form. The most striking disagreement is in the relative rank of the essay examination. In the earlier study only 0.6 per cent of the questions were of the essay type, while in the more recent study 16 per cent of the teachers appear to be using that type extensively. This apparent revival of interest in the essay examination is probably less marked than the difference in ranks between the two studies would indicate, since the earlier tests were written for prize competition. In fact, Lee and Segel² conclude that there was a definite shift toward objective tests.

Davis and Hensley⁴ report that [high school] teachers prefer a combination of essay and objective questions. Fifty-nine per cent of the teachers use a combination of question types. Of the total group four per cent use the essay type question exclusively and thirty-eight per cent use objective examinations only.

Comparative validity and reliability of various types of tests. Ruch⁵ summarized the experimental studies available in 1920 and came

² J. Murray Lee and David Segel, *Testing Practices of High School Teachers*, pages 6-12. United States Office of Education Bulletin No. 9, 1936.

³ J. Murray Lee and David Segel, *op cit.* page 6.

⁴ Iven H. Hensley and Robert A. Davis, "What High-School Teachers Think and Do About Their Examinations," *Educational Administration and Supervision* 38, 219-228, April 1950. Page 220.

⁵ G. M. Ruch, *op cit.* pages 281-306.

to the conclusion that 'the new-type tests are at least as valid as the essay examinations' and that the various objective types are 'not greatly unequal in validity.' Ruch also concluded that 'for equal working times recall and recognition types are not greatly dissimilar although recall tests tended to rank at the top and true-false at the bottom in most of the studies.'

During the next ten years several experimental studies and excellent summaries of the literature were published. Those by Kinney and Eurih⁶ and by Lee and Symonds⁷ were the most comprehensive. The latter study points out that the problem of determination of the comparative merits of different measuring instruments is not only 'one of the most important' it is also 'one of the most poorly done.'

Rinsland⁸ also summarized the experimental literature to 1938 and suggested two cautious conclusions:

1. One might conclude that the objective tests with probably the exception of the true-false type are as valid as or perhaps slightly more valid than the essay or subjective examination and that of all the objective forms the completion or simple recall seems to be the most valid.

2. Generally speaking the various types of objective tests have about equal reliability when compared on the basis of working time. Differences of reliability may be due primarily to the wording of individual items rather than to the objective form.

Lindquist⁹ takes the position that many of the studies which have attempted to determine the comparative validities and reliabilities of various test forms have been 'inconclusive if not definitely misleading.' He points out that these comparative studies have not always recognized the specific nature of test validity, have overemphasized the importance of reliability and have often failed to control such factors as relative skill in constructing the various test forms and the time allotments for the tests. In view of these limitations Lindquist comes to the conclusion that 'in making a selection from a number of test techniques in any specific test situation or in relation to any specific objective of instruction the test constructor must at present depend almost entirely upon logical considerations rather than upon the experimental or empirical evidence that is now available.'

A summary published in 1950¹⁰ concludes that 'few dependable gen-

⁶ L. B. Kinney and A. C. Eurih. *A Summary of Investigations Comparing Different Types of Tests*. *School and Society* 36: 540-544. October 22, 1932.

⁷ J. Murray Lee and Percival M. Symonds. *New Type of Objective Tests: A Summary of Recent Investigations*. *Journal of Educational Psychology* 24: 21-39. February 1933. 25: 161-184. March 1934.

⁸ Henry Daniel Rinsland. *Constructing Tests and Grading in Elementary and High School Subjects*. pages 235-299. New York: Prentice-Hall Inc. 1938.

⁹ Herbert E. Hawkes, E. F. Lindquist and C. R. Mann. *The Construction and Use of Achievement Examinations*. pages 97-103. Boston: Houghton Mifflin Company. 1936.

¹⁰ Max D. Engelhart. *Examinations*. in the *Encyclopedia of Educational Research*, edited by Walter S. Monroe. pages 407-412. New York: The Macmillan Company. 1950.

eralizations can be drawn from studies in this area" and that the "overlappings are much more significant than minor differences between averages" Another article in the same volume¹¹ arrives at this conclusion

The relative effectiveness of a test technique is specific rather than general. It is probable that the validity of a test technique in most fields is more a function of the ingenuity with which it is applied than it is of the test technique employed.

Adequate comparisons between test techniques can therefore be made only for specific material when results are used for a specific purpose when items are constructed with specific insight and ability, and on the basis of validity coefficients computed for equal amounts of testing time when each test is administered at its optimum rate.

To be of practical guidance to the classroom teacher, research should seek answers to such specific questions as the following. In the measurement of what specific objectives in science is the true-false technique of most worth? What testing technique is most effective for measuring vocabulary in a foreign language? What distinctive value, if any, has the rearrangement test in history? For the present, one's choice of the tools of science must depend chiefly upon one's personal judgment and general educational philosophy rather than upon direct experimental evidence.

It is well to recognize that knowledge may exist and function on at least four different levels. The lowest level involves mere *recognition*. A person's general reading vocabulary, as distinguished from his speaking and writing vocabulary, is an example of knowledge where the ability to recognize is the important thing. The next higher level involves *recall*. For knowledge of many types to have value, one must be able to recall it when needed. Familiar examples are one's speaking and writing vocabulary, the names and faces of acquaintances, and the ordinary number combinations in arithmetic. Sometimes one needs to recall separate facts or isolated bits of knowledge, but at other times the organization is important. The person who is an entertaining conversationalist, an interesting letter writer, or an effective public speaker must be able to present his knowledge in a connected form. A still higher level of knowledge involves the ability to *interpret and evaluate*. At this level the learner must have a sufficient understanding of the material to be able to see it in its relationships to other things. The exercise of discrimination and judgment is implied. The highest level of all involves *application*. The person who is able to utilize information acquired in one situation and who applies it to the intelligent solution of problems in a new setting has arrived at true mastery.

It seems reasonable to assume that the type of test used must be appropriate to the level of knowledge being measured. Tests of the multiple-choice and matching types appear adequate for the first level of knowledge.

¹¹ Walter W. Cook, Achievement Tests in the 1930 *Encyclopedia of Educational Research*, page 1468.

Recall tests may be required for the other three levels. Wherever organization is important, the essay type is perhaps more appropriate than the simple recall. However, far more important than the *type* of test is the skill with which it is used. Understanding, evaluation, application, and many other aspects of thinking can be measured by recognition tests, but to do so requires a degree of skill that the regular classroom teacher rarely attains.¹² It is also probably true that most recall tests measure memory only.

B. Simple-Recall Tests

Definition. The *simple recall* test is here somewhat arbitrarily defined as one in which each item appears as a direct question, a stimulus word or phrase, or a specific direction. The response must be *recalled* by the pupil from his past experience rather than merely *identified* from a list of suggested answers supplied by the teacher. The simple-recall test is differentiated from the essay examination primarily upon the basis of length of response required: the typical response to the simple-recall item is short, preferably a single word or phrase. Thus it is sometimes called a short-answer objective test.

Advantages and limitations. This type of test has the obvious advantage of familiarity and "naturalness." It may stimulate desirable study practices and almost completely eliminate guessing as a factor for measurement, thus avoiding two of the most common faults of objective tests. The *simple recall* test is particularly valuable in mathematics and the physical sciences, where the stimulus appears in the form of a problem requiring computation. It also has wider application to test situations presented in the form of maps, charts, and diagrams in which the pupil is required to supply, in spaces provided, the names of parts keyed by numbers or letters.

One limitation of the simple-recall test is that it tends to measure highly factual knowledge, consisting of isolated bits of information. Also the scoring is somewhat laborious and not always entirely objective. These limitations need not be very serious when the tests are carefully prepared, as can be seen from the illustrations which follow.

I Illustrations of Simple Recall Tests

Below are a few sample test items of the simple-recall form that have been taken from standard tests.¹³ Excellent examples of this and other test

¹² For a comprehensive discussion of this problem see William A. Brownell and Committee. *The Measurement of Understanding*. *Forty-Fifth Yearbook of the National Society for the Study of Education*. Part I. 338 pages. Chicago: University of Chicago Press, 1946.

¹³ In the examples of the various types of objective tests that follow an effort has been made to illustrate a wide variety of mechanical arrangements of items as well as of subject matter. It is recognized that they are not all of equal merit. Some of the tests referred to are perhaps out of print.

forms used in a variety of school subjects on all educational levels are to be found in Rinsland ¹⁴

Stone Reasoning Tests in Arithmetic¹⁵

1. James had 5 cents He earned 13 cents more and then bought a top for 10 cents How much money did he have left? Answer _____
- 2 How many oranges can I buy for 35 cents when oranges cost 7 cents each? Answer _____

Sones-Harry High School Achievement Test, Part II¹⁶

- 1 What instrument was designed to draw a circle? (_____)1
- 2 Write "25% of ' as "a decimal times " (_____)2
- 3 Write in figures one thousand seven and four hundredths (_____)3

Cooperative General Mathematics Tests for College Students, Form 1934¹⁷

- 28 How many axes of symmetry does an equilateral triangle have? (_____)
- 29 Eight is what per cent of 64? (_____)
- 30 Write an expression that exceeds M by X (_____)
- 31 Solve the formula $V = \frac{Bh}{3}$ for h (_____)

Iowa Placement Examinations, Chemistry-Training¹⁸

- 1 The atomic weight of K is 39, of Cl , 35.5, of O , 16
What is the molecular weight of $KClO_3$?
- 2 If 7 gm of iron unite with 4 gm of sulphur, how many gm of iron sulphide will be produced? - - - - -

Tests on Everyday Problems in Science, Unit XII¹⁹

- What device is used in a vacuum-cleaner to pump air into the dust bag? (15) - - - - -
- What is the pressure in pounds of ordinary air per square inch? (16) - - - - -

An Exercise from a Biology Workbook²⁰

DIRECTIONS As you locate each part using a hand lens on an actual specimen,

¹⁴ Henry Daniel Rinsland *op cit*, pages 23-222

¹⁵ Devised by C W Stone, and published by Bureau of Publications Teachers College Columbia University

¹⁶ Devised by W W D Sones and David P Harry, Jr, and published by World Book Company

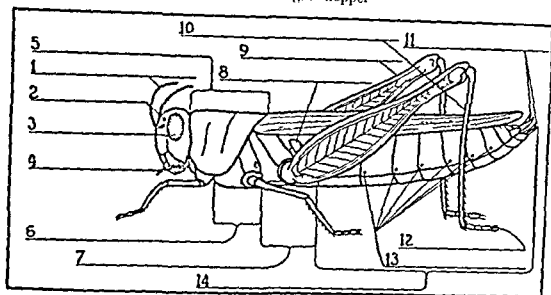
¹⁷ Devised by H T Lundholm and L P Siceloff, and published by Cooperative Test Service

¹⁸ Devised by G D Stoddard and J Cornog and published by Extension Division, State University of Iowa

¹⁹ Devised by C J Pieper and W L Beauchamp, and published by Scott Foresman & Company

²⁰ Prepared by Arthur O Baker and Lewis H Mills to accompany their *Dynamic Biology Today* Chicago Rand McNally & Company, 1943

find the corresponding part in the accompanying illustration and label it. Consider how each part functions in the life of the grasshopper.



Parts of the Grasshopper

The following items from an informal class test illustrate the possibilities of recall tests with more than one response to each item.

For each event below give the country, year, and person with whom you associate it.

Event	Country	Year	Person
First psychological laboratory	_____	_____	_____
First general intelligence test	_____	_____	_____
First standardized achievement test	_____	_____	_____

II Rules and Suggestions for Construction

The simple-recall is one of the most familiar test forms and one of the easiest to prepare. The main problem is how to phrase the test situations so that they will call forth responses of a higher intellectual level than mere rote memory, and so that they can be scored with a minimum expenditure of time and effort.

1. *The direct-question form is usually preferable to the statement form. It is more natural for the pupil and is likely to be easier to phrase.*

EXAMPLE The first president of the United States was
BETTER Who was the first president of the United States?

2. *The questions should be so worded that the response required is as brief as possible, preferably a single word, number, symbol, or at most a short phrase. This will objectify and facilitate scoring.*

3. *The blanks provided for the responses should be in a column, preferably at the right of the questions. This arrangement facilitates scoring and is more convenient for the pupil. The illustrations above show various ways of arranging the answer column.*

forms used in a variety of school subjects on all educational levels are to be found in Rinsland ¹⁴

Stone Reasoning Tests in Arithmetic¹⁵

- 1 James had 5 cents He earned 13 cents more and then bought a top for 10 cents How much money did he have left? Answer _____
- 2 How many oranges can I buy for 35 cents when oranges cost 7 cents each? Answer _____

Sones-Harry High School Achievement Test, Part II¹⁶

- 1 What instrument was designed to draw a circle? (_____)1
- 2 Write 25% of as "a decimal times " (_____)2
- 3 Write in figures one thousand seven and four hundredths (_____)3

Cooperative General Mathematics Tests for College Students, Form 1934¹⁷

- 28 How many axes of symmetry does an equilateral triangle have? (_____)
- 29 Eight is what per cent of 64? (_____)
- 30 Write an expression that exceeds M by X (_____)
- 31 Solve the formula $V = \frac{Bh}{3}$ for h (_____)

Iowa Placement Examinations, Chemistry Training¹⁸

- 1 The atomic weight of K is 39, of Cl , 35.5, of O , 16
What is the molecular weight of $KClO_4$? --
- 2 If 7 gm of iron unite with 4 gm of sulphur, how many gm of iron sulphide will be produced? -----

Tests on Everyday Problems in Science, Unit XII¹⁹

- What device is used in a vacuum-cleaner to pump air into the dust bag? (15) --
- What is the pressure in pounds of ordinary air per square inch? (16) --

An Exercise from a Biology Workbook²⁰

DIRECTIONS As you locate each part using a hand lens on an actual specimen,

¹⁴ Henry Daniel Rinsland *op cit* pages 23-222

¹⁵ Devised by C W Stone and published by Bureau of Publications Teachers College Columbia University

¹⁶ Devised by W W D Sones and David P Harry Jr and published by World Book Company

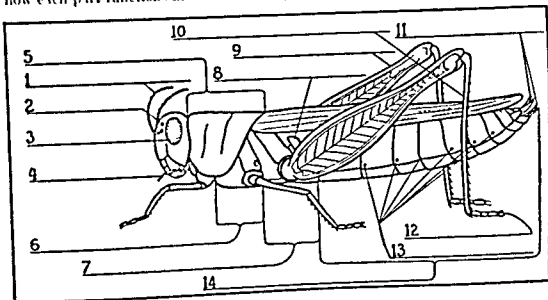
¹⁷ Devised by H T Lundholm and L P Siceloff and published by Cooperative Test Service

¹⁸ Devised by G D Stoddard and J Cornog and published by Extension Division State University of Iowa

¹⁹ Devised by C J Pieper and W L Beauchamp and published by Scott Foresman & Company

²⁰ Prepared by Arthur O Baker and Lewis H Mills to accompany their *Dynamic Biology Text* Chicago Rand McNally & Company, 1943

find the corresponding part in the accompanying illustration and label it. Consider how each part functions in the life of the grasshopper.



Parts of the Grasshopper

The following items from an informal class test illustrate the possibilities of recall tests with more than one response to each item.

For each event below give the country, year, and person with which it is associated.

Event	Country	Year	Person
First psychological laboratory	_____	_____	_____
First general intelligence test	_____	_____	_____
First standardized achievement test	_____	_____	_____

4 *The use of textbook language in wording the question should be reduced to the minimum* Unfamiliar phrasing will reduce the possibility of correct responses that represent mere meaningless verbal associations and also will eliminate the temptation of pupils to memorize the exact language of the book.

5 *The questions should be so worded that there is only one correct response* This is a standard which is difficult to reach since pupils are marvelously resourceful in reading into questions interpretations which the teacher never intended. For example, the question in ancient history, "Name two ancient sports" elicited this reply from an ingenious student "Antony and Cleopatra." This possibility would not have arisen had the question taken this form "What were two popular athletic contests in ancient Greece?" All acceptable replies which are based on any legitimate interpretation of the question should receive credit, and must be listed on the scoring key. Extra care in wording the questions will save much time and trouble later.

C. Completion Tests

Definition. The *completion* test may be defined as a series of sentences in which certain important words or phrases have been omitted and blanks submitted for the pupil to fill in. A sentence may contain one or more blanks. The sentences in the test may be disconnected, or they may be organized into a paragraph. Each blank counts one point.

Advantages and limitations. The mental processes which the pupil must employ in supplying the responses required in completion tests are very similar to those required in simple recall tests, although perhaps on a somewhat higher level. It is not surprising that the advantages and limitations of these two types of tests are also similar. The completion test has wide applicability, as far as subject-matter is concerned, but unless prepared with extreme care is likely to measure rote memory rather than real understanding, or it may turn out to be more a measure of general intelligence or linguistic aptitude than of school achievement.

The scoring is likely to be even more laborious than that of simple-recall tests. This is not only because the scoring is somewhat subjective, but also because the missing words are written in blanks scattered all over the page, rather than in a column. While these limitations cannot be entirely eliminated they can be greatly reduced, as is evident from the illustrations below.

I Illustrations of Completion Tests

Stanford Achievement Test, Paragraph Meaning, 1940 Edition²¹

DIRECTIONS [Abridged] Write JUST ONE WORD on each line. Be sure to write each answer on the line that has the same number as the missing word in the paragraph.

²¹ Devised by Truman L. Kelley, Giles M. Ruch and Lewis M. Terman and published by World Book Company.

123

In olden days men made their own pens from the quills of feathers. It required considerable skill to cut a pen properly so as to suit one's individual taste in writing. Students were always on the lookout for good goose, swan, turkey, or other bird feathers. Goose quills made the most satisfactory —1— for general —2—, but schoolmasters liked pens made from the —3— of swan feathers because they fitted best behind the ear.

Answer

- 1 _____
2 _____
3 _____

Public School Attainment Tests for High School Entrance²¹

3 Question Did this team have a coach?

Answer No they taught (3) how to play without any coach

(3) _____

1 Q Did all of you have matches?

A Of course! Each one had (1) own water proof box full

(4) _____

Tests of Everyday Problems in Science, Unit XI²²

A pry-pole is an example of a machine called the

(11) _____

A capstan is an example of a machine called the

(12) _____

A screw is an example of a machine called the

(13) _____

Your teeth are examples of machines called

(14) _____

Category Tests in American History²³

2 The man who headed the first expedition to circumnavigate the globe was

7 The Articles of Confederation were in force from 1781 to

9 The 'Old Liberty Bell' rang out the decision of Congress to be free from England in the year

Write your words and dates here

- 2 _____
7 _____
9 _____

Cooperative English Test Series I²⁴

20 Write on the lines to the right the contractions—shortened forms to represent how the words are naturally spoken—for the seven groups of words underlined in the following sentences. For instance, for *do not* you would write *don't*. You need not copy the sentence but only the seven contractions.

I have read his story, but *I* cannot believe that *he* will get a passing grade on it for it is not well written and *has* not a clear cut plot. The characters are not at all interesting *they* are not even human.

- _____
- _____
- _____
- _____
- _____

²¹ Devised by Henry D. Rinsland and Roland L. Beck and published by Public School Publishing Company. (The form of completion with all responses in a column instead of staggered within sentences was devised by Rinsland.)

²² Devised by C. J. Pieper and W. L. Beauchamp and published by Scott, Foresman & Company.

²³ Devised by C. A. Gregory and published by C. A. Gregory Company.

²⁴ Devised by Sterling A. Leonard and others and published by Cooperative Test Service, 1931.

II Rules and Suggestions for Construction

Most of the suggestions made for constructing simple-recall tests apply equally well to completion tests. The dangers to be avoided are largely the same for both forms. A few suggestions may be offered, however, that have special reference to completion items. The main problems of constructing completion tests are three in number: (1) How to phrase the statements so as to indicate the *type* of response desired, (2) How to avoid giving the pupil unwarranted clues to the specific responses expected, and (3) How to arrange the items so as to facilitate scoring. The first two suggestions below apply to problem one, the next five apply to problem two, and the last five suggestions are for problem three. In short, a good completion statement gives a *reasonable basis* for determining the response desired without providing *unwarranted clues*, and is arranged to *facilitate scoring*.

1 *Avoid indefinite statements.* The pupil is entitled to know the *type* of response desired, and when this is done the scoring is far more rapid.

EXAMPLE Abraham Lincoln was born in _____

BETTER Abraham Lincoln was born in the state of _____
The year of Abraham Lincoln's birth was _____

The first statement fails to indicate whether the desired response is the date, the place, or the circumstances of his birth. In that form legitimate answers might be "February" or "1809," for the date, "Kentucky," or possibly "The South," for the place, and "poverty," or "a log cabin" for the circumstances of his birth. By a slight change in wording the statement is made quite definite.

2 *Avoid overmutilated statements.* If too many key words are left out, it is impossible to know what meaning was intended.

EXAMPLE The _____ is obtained by dividing the _____ by the _____.

In its present form, it is impossible to tell whether the statement refers to educational measurement or to arithmetic.

BETTER

1 The IQ is obtained by dividing the _____ by the _____.

2 The _____ is obtained by dividing the _____ by the divisor _____.

3 *Omit key words and phrases, rather than trivial details.* If this is not done the response may be as obvious as the first example below, or as unnecessarily difficult as the second example.

EXAMPLES

1 Abraham Lincoln was born in February _____ 1809.

2 Abraham Lincoln was born in _____ County, Kentucky.

4 *Avoid stating statements directly from the text.* This puts too great a premium on rote memory.

5 *Whenever possible avoid 'a' or 'an' immediately before a blank.* These words unnecessarily limit the responses that can be used in the blank.

EXAMPLE Mary picked an _____ off the tree and ate it

BETTER Mary ate the _____ which she picked off the tree

It is apparent that the words "pear," "peach," "plum," "cherry," "lemon," "pineapple," and the like could not be used in the first statement. In fact, the choice tends to narrow down to two familiar fruits, "apple," or "orange." The second statement contains no specific determiner.

6 *Make the blanks of uniform length.* If the blanks vary in length the pupil has a clue to the length of answer expected. Even more of a clue is afforded by using a dot or a dash for each letter in the correct word.

EXAMPLE

1 The second president of the United States was _____ from the state of _____

2 The president in office during the Mexican War was _____ from the state of _____

BETTER

1 The second president of the United States was _____ from the state of _____

2 The president in office during the Mexican War was _____ from the state of _____

7 *Atoid grammatical clues to the answer expected*

EXAMPLE The authors of the first performance test of intelligence were _____

BETTER The first performance test of intelligence was prepared by _____

8 *Choose statements in which there is only one correct response for the blanks.* The scoring is far more objective if only one specific word or phrase can be used to complete the statement.

9 *The required response should be a single word or a brief phrase.* The more the scorer has to read the more time will be required.

10 *Arrange the test so that the answers are in a column at the right of the sentences.* The illustrations above show various ways in which this may be done. When each sentence contains but a single blank, the scoring is made easier if the blank comes at the end. The Tests of Everyday Problems in Science and the Gregory Tests in American History are examples. If the sentences contain more than one blank the scoring is more rapid if the blanks are numbered and the pupil is directed to write his responses in the correspondingly numbered blank in the answer column at the right. Rinsland²⁶ suggests that the following wording of the directions will be clear to pupils above the fourth grade although it may be necessary to explain the word "correspondingly" in grades four to seven.

DIRECTIONS In each of the sentences below, one or more words, numbers or dates are needed in the numbered blank spaces to make the sentences complete and true. Place the word or words in the correspondingly numbered blank to the right.

11 *Prepare a scoring key which contains all acceptable answers.* Although it is desirable to have only one response which can be considered correct for each blank,

²⁶ Henry Daniel Rinsland *op cit*, page 56

this is not possible in all cases. As a rule, a satisfactory key can be made by writing in red the correct answers on a copy of the test.

12 Allow one point for each blank correctly filled. Avoid fractional credits and unequal weighting of items on the basis of difficulty or importance.

D. Alternative-Response Tests

Definition. An *alternative-response* test is made up of items each of which admits of only two possible responses. The usual form is the familiar true-false test. Other similar forms are right-wrong, correct-incorrect, yes-no, same-opposite, and two-option multiple-choice.

Advantages and limitations. Obvious advantages of the alternative-response test are its apparent ease of construction, applicability to a wide range of subject-matter, objectivity of scoring, and wide sampling of knowledge tested per unit of working time. The true-false test, a form very popular with classroom teachers, has been the object of more research and of more criticism than any other form of objective test. The negative-suggestion effect and the factor of guessing are often pointed out as limitations of this type of test. While the use of the correction formula appears to make a fairly satisfactory adjustment for guessing in the total score, the alternative-response is not well adapted to educational diagnosis. The danger of negative suggestion when pupils see statements which are false has apparently been overestimated, but perhaps it is wise not to use true-false tests as pretests or with young children. In such cases it is better to avoid the alternative-response test, or to use a question that can be answered by *yes* or *no* instead of a declarative statement.

Several modifications and alleged improvements of the true-false test have been proposed. Barton,²⁷ for example, has suggested crossing out the part of the statement that is in error, while other studies²⁸ have shown that having pupils correct the wrong statements increases the reliability of the test. Still others²⁹ have proposed that items be weighted according to the judgment of the pupil, or be marked *true*, *false*, *doubtful*. All of these suggestions add somewhat to the labor of scoring and have not received wide acceptance. Furthermore, strictly speaking, when these modifications are followed, the test is no longer of the alternative-response type. As a rule, the most obvious way to "improve" the true-false test is also the best,

²⁷ W. A. Barton, Jr., "Improving the True-False Examination," *School and Society* 34: 544-546, October 17, 1931.

²⁸ Ernest E. Bayless and Ralph C. Bedell, "A Study of Comparative Validity as Shown by a Group of Objective Tests," *Journal of Educational Research* 23: 8-16, January, 1931; F. D. Curtis, W. C. Darling, and N. H. Sherman, "A Study of the Relative Values of Two Modifications of the True-False Test," *Journal of Educational Research* 36: 517-527, March, 1943; W. H. E. Wright, "The Modified True-False Item Applied to Testing in Chemistry," *School Science and Mathematics*, 44: 637-639, October, 1944.

²⁹ Kate Hevner, "A Method for Correcting for Guessing in True-False Tests and Empirical Evidence in Support of It," *Journal of Social Psychology*, 3: 359-362, August 1932.

that is, *make the test longer and prepare it more carefully*. At least 75 items are desirable, and 50 may be set as an absolute minimum, unless the test covers a very narrow range or is used for instructional purposes only. One advantage of the true-false test is that it can cover more items in the same time than any other test type.

Should pupils be advised to look over true-false tests and change the answers on doubtful items? Several studies have attempted to answer this question. Hill³⁰ made an extensive investigation of the problem and came to the conclusion that there is "not much advantage to be gained by changing one's answers on a true-false test," although the advantage was somewhat greater in changing from true to false than in the reverse. There is some evidence that the better pupils profit most from rechecking and revising their work. Even if the scores are not always improved, it is probably a good work habit to encourage.

The low esteem in which test experts hold the alternative-response type of test, especially the true-false form, is indicated by the infrequency with which it has appeared in recent standardized achievement tests. This is due chiefly to its weakness as an instrument of diagnosis and to the fact that such tests must be made much longer than other objective tests in order to secure comparable reliability. Although this type of test has been greatly overworked by classroom teachers, it does have a legitimate, though restricted, use in informal tests. For example, the true-false test seems well adapted to testing the persistence of popular misconceptions and superstitions. Ordinary alternative-test situations are encountered in which it is difficult or impossible to make more than two plausible responses for a multiple-choice test. There are many troublesome situations of this sort in language usage. Common examples include the case forms in pronouns, correct use of singular and plural verbs, confusions of past tense and past participles, the use of *sit* and *set*, *lay* and *lie*, and many others. A safe rule would be to *restrict the use of the alternative-response test to those situations to which other test forms are inapplicable, and then to give particular care to the wording of the items*.

1 Illustrations of Alternative Response Tests

California Achievement Tests—Advanced Battery Form 4A³¹

DIRECTIONS In the following sentences mark as you have been told the number of each correct word.

Test 5—Section C

- | | |
|---|----------|
| 36 (Isn't 'Aren't) the baskets filled with flowers? | _____ 36 |
| 47 I approve of (his 'him) going | _____ 47 |

³⁰ George E. Hill, "The Effect of Changed Responses in True-False Tests," *Journal of Educational Psychology* 28, 308-310, April 1937.

³¹ Devised by Ernest W. Tieg and Willis W. Clark and published by California Test Bureau.

For each statement given below that is a complete sentence, mark YES, for each that is not mark NO

- | | | | | |
|----|--|-----|----|----|
| 51 | When we approached the deserted farmhouse at night | YES | NO | 51 |
| 56 | The mountains resounded with peals of thunder which indicated the storm's fury | YES | NO | 56 |

Iowa Silent Reading Tests, New Edition Sentence Meaning
Elementary, Form Am²²

DIRECTIONS Read each question. If the answer is "Yes," fill in the space under YES in the margin. If the answer is "No," fill in the space under NO. Study the sample. Do not guess.

- | | | | | |
|---|---|---|-----|----|
| 1 | Is a dime less in value than a nickel? | 1 | YES | NO |
| 2 | Can we see things clearly in a thick fog? | 2 | YES | NO |
| 3 | Is geography studied in public schools? | 3 | YES | NO |

Allport-Vernon Lindzey 'Study of Values'²³

DIRECTIONS A number of controversial statements or questions with two alternative answers are given below. Indicate your personal preferences by writing appropriate figures in the boxes to the right of each question. For each question you have three points that you may distribute in any of the following combinations:

If you agree with alternative (a) and disagree with (b), write [3 in the first box and 0 in the second box]

If you agree with (b), disagree with (a), write [0 in the first box and 3 in the second box]

If you have a slight preference for (a) over (b), write [2 in the first box and 1 in the second box]

If you have a slight preference for (b) over (a), write [1 in the first box and 2 in the second box]

- 1 The main object of scientific research should be the discovery of truth rather than its practical applications (a) Yes, (b) No

a	b
<input type="text"/>	<input type="text"/>

- 10 If you were a university professor and had the necessary ability, would you prefer to teach (a) poetry, (b) chemistry and physics?

a	b
<input type="text"/>	<input type="text"/>

Tests in English Fundamentals Grammar²⁴

DIRECTIONS Classify the italicized words in the sentences below as adjectives or adverbs by placing check marks in the proper columns

²² Devised by H. A. Greene and V. H. Kelley and published by World Book Company.

²³ Devised by Gordon W. Allport, Philip E. Vernon, and Gardner Lindzey and published by Houghton Mifflin Company, 1951.

²⁴ Devised by R. Davis and published by Ginn and Company.

- 3 That was a *silly* remark
- 6 Those flowers smell *sweet*
- 11 You can *hardly* expect him to wait

	Adjective	Adverb
3		
6		
11		

The Iowa Every Pupil Tests in Basic Skills²⁵

DIRECTIONS In each of the following sentences there are two or more numbered words or phrases inclosed in brackets. If you think the *first* word or phrase is correct place an X in the *first* box of the corresponding row on the answer sheet. If you think the *second* answer is correct place an X in the *second* box of the proper row, etc.

- 7 Ted is $\left\{ \begin{array}{l} 1 \text{ a} \\ 2 \text{ an} \end{array} \right\}$ industrious man
- 54 My father $\left\{ \begin{array}{l} 1 \text{ has} \\ 2 \text{ hasn't} \end{array} \right\}$ no money
- 62 I want everyone to help $\left\{ \begin{array}{l} 1 \text{ himself} \\ 2 \text{ themselves} \end{array} \right\}$

Cooperative Plane Geometry Test Revised Series Q²⁶

DIRECTIONS Read these statements and mark each one in the parentheses at the right with a plus sign (+) if you think it is always true or with a zero (0) if you think it is always or sometimes false.

- 1 The opposite angles of a parallelogram are equal 1()
- 17 If two triangles are similar their areas are in the same ratio as the medians drawn to corresponding sides 17()
- 2 A diameter of a circle divides the circle into two equal parts 2()
- 18 All similar polygons are equilateral 18()

Tests on Everyday Problems in Science Unit III²⁷

DIRECTIONS There are 25 incomplete statements in this test each followed by parts (a) (b) (c) and (d). One or more of these parts or perhaps none of them correctly complete the incomplete statement. You are to place a plus sign (+) in the parentheses (near the right margin) opposite each part which correctly completes the statement and a minus sign (-) opposite each part which does not correctly complete the statement.

²⁵ Devised by H. A. Gree and published by Extension Division, State University of Iowa, 1939.

²⁶ Devised by Emma Slattery and L. P. Siceloff and published by the Cooperative Test Service.

²⁷ Devised by C. J. Harper and W. L. Beauchamp and published by Scott Foresman & Company.

- 13 Minerals in our food supply ()
 (a) furnish heat and energy to the body ()
 (b) are the only materials of which cells can be built ()
 (c) are good regulators of certain of the body activities ()
 (d) help particularly to build bone and blood ()

Cooperative Solid Geometry Tests²³

DIRECTIONS Read these statements and mark each one in the parentheses at the right with a plus sign (+) if you think it is true, or with a zero (0) if you think it is false, wholly or in part

- 4 Any number of planes may be passed through a given straight line ()
 27 Two planes parallel to the same straight line are parallel to each other ()
 41 The square of a diagonal of a cube is three times the square of its edge ()

George Washington University English Literature Test²⁴

- T F 1 "Il Penseroso" describes the charms of a merry social life
 T F 4 "Pilgrim's Progress" is one of the greatest prose allegories in literature
 T F 8 In his poem "The Bells," Poe describes the process of making bells

II Rules and Suggestions for Construction

The true-false test is often thought to be one of the easiest types to prepare. This superiority is more apparent than real, however. Experienced test makers are convinced that no test form demands greater skill. Unusual care must be exercised in wording true-false statements so that the *content* rather than the *form* of the statement will determine the response. The aim should be to phrase the statement so as not to make its meaning needlessly obscure on the one hand, nor to provide unwarranted clues on the other. This balance requires a delicate skill of adjustment that is rare among makers of informal tests. The following specific suggestions may be found helpful in constructing true-false tests. Many of the suggestions for constructing multiple-choice tests that are found in the next section are also applicable here.

1 *Avoid specific determiners.* It has been found that strongly worded statements are much more likely to be false than true, while moderately worded statements are much more likely to be true than false. Examples of the former are those containing "all," "always," "never," "no," "none," "nothing," and the like, examples of the latter are those containing "may," "some," "sometimes," "often," "as a rule," and the like. If care is taken to balance the proportion of true and false items containing any particular expression, that expression ceases to be a specific determiner that affords a clue to the answer.

2 *Avoid a disproportionate number of either true or false statements.* Since several studies have shown that false statements are more valid than true statements²⁵ the suggestion is sometimes made that the test should have more false statements

²³ Devised by H. T. Lundholm and others and published by Cooperative Test Service, 1934.

²⁴ Devised by K. T. Omwake and others, and published by Center for Psychological Service.

than true. If this were generally done, however, the validity of the false statements would probably be reduced, since the pupil would then tend to mark all doubtful statements false. Tell the students that *approximately* half of the statements are true, the other half false.

3 *Avoid the exact language of the textbook.* Lifting true statements directly from the textbook, or making false statements by changing a single word or expression puts too great a premium on rote memory.

4 *Avoid trick statements.* These are usually statements which appear to be true but which are really false because of some inconspicuous word or phrase.

EXAMPLES 1 "The Raven" was written by Edgar Allen Poe

2 The battle of Hastings was fought in 1066 B.C.

BETTER 1 "The Raven" was written by Edgar Allan Poe

2 The battle of Hastings was fought in 55 B.C.

Also avoid "double-headed" statements like the following one (especially if they are partly true and partly false). Poe wrote "The Gold Bug" and "The Scarlet Letter."

5 *Avoid double negatives.* Such statements are especially bad, since pupils well versed in English grammar might conclude that two negatives equal an affirmative, while other pupils would interpret such statements as emphatic negatives.

6 *Avoid ambiguous statements.* With one interpretation the statement may be true and with another equally plausible interpretation it may be false. It is impossible to tell what is being measured when a statement has more than one legitimate interpretation.

7 *Avoid unfamiliar, figurative, or literary language.* The experience of the learner must be considered. A statement is badly worded when a pupil who understands the point involved misses it because of the language employed.

8 *Avoid long statements, especially those involving complex sentence structure.* Same reason as for the preceding suggestion.

9 *Avoid qualitative language wherever possible.* Quantitative language conveys more exactly the meaning intended. Expressions such as "few," "many," "large," "small," "old," "young," "important," "unimportant" are vague and indefinite.

10 *Require the simplest possible method of indicating the response.* Instead of requiring the pupil to write *True* and *False* or *Yes* and *No*, let him write *T* and *F*, *Y* and *N*, or underline the correct response. The symbols "+" for true and "0" for false are so distinct as to make scoring still easier. When the pupil must choose between two words or expressions, the responses should be numbered so that they can be indicated by writing the correct number.

11 *Indicate by a short line or by () where the response is to be recorded.* The responses may be arranged in a column at either the left or right of the statements. Most scorers prefer the answers at the right.

12 *Arrange the statements in groups.* There is some advantage in scoring if the items are arranged in groups of five, with double spacing between each group.

E. Multiple-Choice Tests

Definition. A multiple-choice test is made up of items each of which presents two or more responses, only one of which is correct or definitely

better than the others⁴⁰ Each item may be in the form of a direct question, an incomplete statement, or a word or phrase This form of test is to be distinguished from the *multiple response* type, which requires that two or more responses be made to a single item

Possibilities and limitations. The multiple-choice type of item is usually regarded as the most valuable and most generally applicable of all test forms Lee regards it as "one of the best means for testing judgment that is available"⁴¹ Lindquist asserts that it is "definitely superior to other types" for measuring such educational objectives as "inferential reasoning, reasoned understanding, or sound judgment and discrimination on the part of the pupil"⁴² Cronbach⁴³ regards it as being practically free from "response sets," the tendency for examinees to select a given option position more often than would be predicted on the basis of chance alone

One study⁴⁴ suggests fourteen types of questions which may be asked in multiple-choice test items The list is not all inclusive and does not intend to prescribe the exact language to be used but serves as a guide in formulating the questions

- 1 Definition
 - a What means the same as ?
 - b What conclusion can be drawn from ?
 - c Which of the following statements expresses this concept in different form?
- 2 Purpose
 - a What purpose is served by ?
 - b What principle is exemplified by ?
 - c Why is this done?
 - d What is the most important reason for ?
- 3 Cause
 - a What is the cause of ?
 - b Under which of the following conditions is this true?
- 4 Effect
 - a What is the effect of ?
 - b If this is done, what will happen?
 - c Which of the following should be done (to achieve a given purpose)?
- 5 Association

What tends to occur in connection (temporal causal, or concomitant association) with ?

⁴⁰ It is also possible especially in English usage and spelling tests to have several correct forms and only one incorrect or least desirable form which is to be chosen in each item.

⁴¹ J. Murray Lee *A Guide to Measurement in Secondary Schools* page 379 New York D Appleton-Century Company 1936

⁴² Herbert E. Hawkes E. F. Lindquist and C. R. Mann, *op cit* page 138

⁴³ Lee J. Cronbach 'Further Evidence on Response Sets and Test Design' *Educational and Psychological Measurement* 10 3-31 Spring 1950

⁴⁴ Charles I. Mosier M. Claire Myers and Helen G. Price 'Suggestions for the Construction of Multiple-Choice Test Items' *Educational and Psychological Measurement* 5 261-271 Autumn 1945

6. Recognition of Error
Which of the following constitutes an error (with respect to a given situation)?
7. Identification of Error
 - a. What kind of error is this?
 - b. What is the name of this error?
 - c. What recognized principle is violated?
8. Evaluation
What is the best evaluation of _____ (for a given purpose) and for what reason?
9. Difference
What is the important difference between _____?
10. Similarity
What is the important similarity between _____?
11. Arrangement
In the proper order (to achieve a given purpose or to follow a given rule) which of the following comes first (or last or follows a given item)?
12. Incomplete Arrangement
In the proper order, which of the following should be inserted here to complete the series?
13. Common Principle
All of the following items except one are related by a common principle
 - a. What is the principle?
 - b. Which item does not belong?
 - c. Which of the following items should be substituted?
14. Controversial Subjects
Although not everyone agrees on the desirability of _____, those who support its desirability do so primarily for the reason that _____.

Unusual care must be exercised in the construction of multiple-choice items in order to avoid the inclusion of irrelevant or superficial clues, and to insure that the tests measure something more than the memory of factual knowledge. The value of multiple-choice tests in diagnosis depends upon the skillful selection of the incorrect choices presented in the items.⁴⁵

I. Illustrations of Multiple-Choice Tests

The items below, taken from standard tests, illustrate several different arrangements of multiple-choice tests in a variety of subjects.⁴⁶ This type of test is widely used in all school subjects, on all educational levels, and for measuring a variety of teaching objectives.

Special attention should perhaps be called to two of the illustrations, both of which are suggestive to teachers in making informal tests. The Nelson High School English Test illustrates the possibility of testing punc-

⁴⁵ Ellis Weitzman and Walter J. McNamara, "Apt Use of the Inept Choice in Multiple-Choice Testing," *Journal of Educational Research* 39: 517-522, March 1946.

⁴⁶ These tests are not all equally good, however. The reader will note that some of them are not wholly consistent with the principles set forth in this chapter.

uation with a minimum of scoring labor The Cooperative Test of Social Studies Abilities shows how objective tests may be used to test more than the memory for factual knowledge This is a good example of a test of the pupil's ability to interpret facts—an ability which is an important aspect of thinking

Kuhlmann Finch Intelligence Tests, Test IV⁴⁷

13 Early is to begin as late is to

1	2	3	4	5	
start	end	awake	enter	prompt	13 ----

22 Flour is to bread as sugar is to

1	2	3	4	5	
sweet	candy	fruit	cook	eat	22 ----

The Modern School Achievement Tests Language Usage⁴⁸

DIRECTIONS In each sentence, choose the word or group of words that make the best sentence Then on the dotted line at the right copy the number that is before the correct form

4 I borrowed a pen 1 off
 2 off of my brother
 3 from

7 Every student must do 1 your
 2 his best
 3 their

17 He 1 has got
 2 has his violin with him
 3 has gotten

The Barrett-Ryan Literature Test Silas Marner⁴⁹

- A () An episode that advances the plot is the—1 murdering of a man 2 kidnapping of a child 3 stealing of money 4 fighting of a duel
B () Dolly Winthrop is—1 an ambitious society woman 2 a frivolous girl 3 a haughty lady 4 a kind helpful neighbor
C () A chief characteristic of the novel is—1 humorous passages 2 portrayal of character 3 historical facts 4 fairy element

Wesley Test in Political Terms⁵⁰

- 1 An embargo is
 1 a law or regulation 2 a kind of boat 3 an explorer
 4 a foolish adventure 5 an embankment ()
2 An injunction is a
 1 part of speech 2 wreck 3 union of two things
 4 court order 5 form of advice ()

⁴⁷ Devised by F H Finch F Kuhlmann and G L Betts and published by Educational Test Bureau

⁴⁸ Devised by A I Gates and others and published by Bureau of Publications Teachers College Columbia University

⁴⁹ Devised by E R Barrett, T M Ryan and H E Schrammel and published by Kansas State Teachers College Emporia

⁵⁰ Devised by E B Wesley and published by Charles Scribner & Sons

Unit Scales of Attainment in Foods and Household Management²²

- 2 The spoon should be placed
 1 at the top of the plate
 2 at the left of the fork
 3 in the spoon holder on the table
 4 at the right of the knife
- 40 We get the most calories per pound from
 1 proteins 2 carbohydrates
 3 fats 4 mineral matter
 5 vitamins

Traxler Silent Reading Test, Word Meaning²³

- 5 The commendation is deserved
 (1) success (2) blow (3) popularity (4) good fortune
 (5) praise
- 9 His actions received condemnation
 (1) approval (2) applause (3) censure (4) sympathy
 (5) contempt

Cooperative French Test, Junior Form 1936²⁴

- 2 Quand on vous pose une question, il faut
 1 répondre, 2 se taire, 3 se sauver, 4 tourner le dos, 5 baisser la tête
- 7 Cette dame est ma grand'mère, je suis
 1 son fils, 2 son neveu, 3 son frère, 4 son cousin, 5 son petit-fils
- 16 J'ai deux frères, Jean et Paul Jean a sept ans Paul en a treize et moi
 j'ai douze ans Qui est le plus jeune?
 1 Jean, 2 Paul 3 moi, 4 dix ans, 5 les deux frères

Nelson High School English Test²⁵

DIRECTIONS Some of the sentences contain errors in punctuation, some of them are correct. If you think some mark is not needed, cross out the letter indicating that mark under the word "Omit." If you think some additional mark is needed, cross out the letter indicating that mark under the word "Add." If you think the exercise is correct, cross out the letter r. Key a—apostrophe, c—comma, d—dash e—exclamation point, h—hyphen, p—period q—quotation mark, s—semicolon

	Add	Omit	Right
1 You must elect a chairman, three judges and an official timekeeper	X h d s	q	r
6 He said ' that either you or I must go '	c s d e	X	r
8 The car which John is driving is a new one	d q s c	d	X
14 "Well I think highly of them Mary ' I said	e h s X	p	r

²² Devised by Ethel B. Reeve and Clara M. Brown and published by Educational Test Bureau, Inc.

²³ Devised by Arthur L. Traxler and published by Public School Publishing Company.

²⁴ Devised by Jacob Greenberg and Geraldine Spaulding and published by Cooperative Test Service.

²⁵ Devised by M. J. Nelson and published by Houghton Mifflin Company.

Cooperative Test of Social Studies Abilities, Experimental Form Q¹⁵

INTERPRETING FACTS

DIRECTIONS The exercises in this part consist of a series of paragraphs each followed by several statements about the paragraph. In the parentheses after each statement, put a

- 1, if the statement is a reasonable interpretation, fully supported by the facts given in the paragraph,
 - 2, if the statement goes beyond and cannot be proved by the facts given in the paragraph,
 - 3, if the statement contradicts the facts given in the paragraph
- [The sample exercise and the explanation are omitted]

I. The nineteenth century witnessed a rapid growth in Germany's industrial power. Like England, Germany came to have a fairly satisfactory balance between the amount of its export and import trade. Heavy exports of coke supplied full cargoes for ships to foreign ports and helped to balance heavy importations of raw materials. The imports especially provided a means for distributing freight rates to the advantage of the German trader competing overseas. By these means Germany was constantly obtaining larger portions of world trade. German wares were carried into every trading realm, and trade meant political as well as commercial power in foreign lands.

1. Through growth in foreign trade, Germany's industrial power increased in the nineteenth century. (1)
2. Germany had an export trade equal in volume to that of England. (2)
3. Germany exported very little coke to foreign countries. (3)
4. England was unable to balance the tonnage of her import and export shipments. (4)
5. By reducing freight rates, Germany was constantly gaining a greater percentage of world trade. (5)
6. The sale of German wares in every part of the world resulted in added political influence and commercial growth. (6)

Several kinds of multiple-choice items, including analogies, are illustrated in Appendix A, pages 429-435.

II Rules and Suggestions for Construction

The purpose of suggestions 1 to 5 below is to avoid unwarranted clues to the desired response, the purpose of suggestions 6 to 9 is to encourage responses on a high intellectual level, and the purpose of suggestions 10 to 14 is to make the scoring as simple and rapid as possible.

1 *Make all optional responses grammatically consistent.* For example, if the verb is singular, avoid plural responses and vice versa. Avoid using "a" or "an" as the word in an incomplete statement immediately preceding the list of responses unless all options begin with consonant sound (in the case of "a") or all begin with a vowel sound (in the case of "an").

2 *As a rule, use direct questions rather than incomplete statements.* The question form is more natural and less likely to contain irrelevant clues.

3 *Avoid making the correct response consistently longer or shorter than the others.*

¹⁵ Devised by J. Wayne Wrightstone and published by Cooperative Test Service.

4 Avoid using in the correct response the same words or phrases that occur in the question or incomplete statement

5 Arrange the responses so that the correct one occurs in random order The pupils are likely to detect any regularly recurring pattern in the sequence of response

6 Make all responses plausible In phrasing multiple-choice test items, consideration should be given to the fact that the answer may be arrived at by eliminating the incorrect responses as well as by selecting the correct response directly The aim should be to make each suggested response so plausible as to tempt pupils who have only superficial knowledge of the point involved The plausibility of incorrect responses may be increased by using familiar, stereotyped or textbook phraseology, or expressions very similar to those in the question or incomplete statement

7 At least four choices should be presented whenever possible Increasing the number of plausible choices tends to reduce the guessing factor Horst¹⁶ found, however, that when the incorrect responses are of equal difficulty the chance element is less than when the choice is among a greater number of responses with a wider range of difficulty

8 In testing for the understanding of a term or concept the term should usually be presented first, followed by a series of definitions or descriptions from which the choice is to be made If the order is reversed so that from a series of terms the choice is made of the one that best fits the definition or descriptive statement the selection frequently can be made based upon superficial verbal associations and not upon genuine understanding

9 To measure the higher levels of understanding increase the homogeneity of the options provided The following illustration from Lindquist¹⁷ shows how the degree of required discrimination increases with the homogeneity of the responses presented

- A Engel's law deals with
 - 1 the coinage of money
 - 2 the inevitableness of socialism
 - 3 diminishing returns
 - 4 marginal utility
 - 5 family expenditures
- B Engel's law deals with family expenditures for
 - 1 luxuries
 - 2 food
 - 3 clothing
 - 4 rent
 - 5 necessities
- C According to Engel's law, family expenditures for food
 - 1 increase in accordance with the size of the family
 - 2 decrease as income increases
 - 3 require a smaller percentage of an increasing income
 - 4 rise in proportion to income
 - 5 vary with the tastes of families

To respond correctly to A all that is required is the knowledge that Engel's law deals with family expenditures In B a knowledge of the specific item of expenditure is necessary The maximum degree of discrimination, however, is required in C where still more information is given

¹⁶ Paul Horst The Difficulty of a Multiple-Choice Test Item *Journal of Educational Psychology* 24 229-232 March 1933

¹⁷ Herbert E Hawkes E F Lindquist and C R Mann *op cit* pages 146-147

10 *Require the simplest possible method of indicating a response* This usually means that the responses are lettered and the choice is made by indicating the letter of the response In the first two or three grades where key letters may not be understood, it will be better to permit the more natural response of underlining the correct answer

11 *Indicate by a short line or 1/2 () where the response is to be recorded or, better still, use a separate answer sheet*

12 *Arrange the items in groups* As a rule, groups of five will be suitable although other numbers of items may sometimes be better Double space between each group

13 *Use the "correction for chance" formula (page 156) if the number of choices is fewer than six* If there are six or more responses suggested for each item the gain in validity is seldom sufficient to warrant the labor of making corrections for chance

14 *Group together all items with the same number of choices* This is especially desirable when the correction formula is to be used

F. Matching Tests

Definition. A *matching* test typically consists of two columns, each item in the first column to be paired with a word or phrase in the second column upon some basis suggested In the simplest form of matching test the number of responses is exactly the same as the number of items Frequently, matching tests are made which provide more responses than are required Sometimes the items in the first column are incomplete sentences, each of which requires a word or phrase from the second column for its completion Occasionally two, or even more, columns of responses are given, from each of which a choice must be made for each item in the first column The matching test is also useful for identifying numbered places or parts on maps, charts, and diagrams

Advantages and limitations. There are many types of learning which involve the association of two things in the mind of the learner Common examples are the following Events and dates events and persons, events and places terms and definitions, foreign words and English equivalents, laws and illustrations rules and examples, tools and their use, and the like The matching test is a very convenient form of exercise for measuring such learning In the words of Lindquist "The matching exercise is particularly well adapted to testing in *who*, *what*, *when*, and *where* types of situations, or for naming and identifying abilities"⁵⁸

Its principal limitations are as follows (1) It is not well adapted to the measurement of understanding as distinguished from mere memory, (2) With the exception of the true-false test, the matching test is the form most likely to include irrelevant clues to the correct response and (3) Unless skillfully made, it is time consuming for the pupil The suggestions that follow are designed to overcome the last two limitations The matching test can hardly be designed to measure genuine understanding of a high level or the ability to interpret complex relationships

⁵⁸ Herbert F. Hawkes, F. F. L. J., and C. J. M.

I Illustrations of Matching Tests

The following examples from standard tests illustrate different mechanical arrangements of matching tests in a variety of subjects

Every Pupil Test in Physics⁴⁹

DIRECTIONS Read each definition or description. Then select from the Answer List the word thus defined and write its number on the dotted line in front of the definition. The answer to the sample is (Power), so 18 is written on the dotted line.

ANSWER LIST (Arranged alphabetically)

- | | | |
|---------------|-------------------------|----------------------|
| 1 Adhesion | 10 Energy | 17 Potential |
| 2 Centrifugal | 11 Heat of Fusion | 18 Power |
| 3 Centripetal | 12 Heat of Vaporization | 19 Radiation |
| 4 Cohesion | 13 Inertia | 20 Relative Humidity |
| 5 Conduction | 14 Insulator | 21 Specific Gravity |
| 6 Conductor | 15 Kinetic | 22 Specific Heat |
| 7 Convection | 16 Mechanical Advantage | 23 Surface Tension |
| 8 Density | | 24 Work |
| 9 Efficiency | | |

18 SAMPLE The rate of doing work

- 1 Weight per unit volume
- 2 Mutual force of attraction between like molecules
- 3 Tendency of a body to resist any change in its state of rest or motion
- 4 Tendency of surface of a liquid to contract as much as possible
- 5 Capacity for doing work
- 6 The ratio of resistance overcome to effort exerted
- 7 The product of a force and the distance through which it acts
- 8 Ratio of output to input
- 9 The energy a body possesses because of its position
- 10 The number of calories required to melt one gram of a substance
- 11 Amount of water-vapor the air holds compared to what it could hold at the same temperature
- 12 Transfer of heat from a hot to a cold body by molecular collision
- 13 Transfer of heat by means of ether waves
- 14 The force pulling the body toward the centre of rotation
- 15 A substance that conducts heat or electricity very poorly or almost not at all

Cooperative Test of Social Studies Abilities Experimental Form Q⁵⁰

DIRECTIONS In which of the sources listed in the left-hand column would you look first to find the items listed in the right-hand column? Consider each group separately. Put the number of the best source in the parentheses after each item.

- | | | |
|--------------------------|--|-------|
| 1 Atlas | 51. A discussion of an important present-day issue in Congress | 51() |
| 2 <i>Current History</i> | 52. The location of the ten largest cities in the world | 52() |
| 3 Dictionary | | |
| 4 Economics textbook | | |

⁴⁹ Devised by F. W. Brown and others and published by the Ohio State Department of Education, 1930.

⁵⁰ Devised by J. Wayne Wrightstone and published by Cooperative Test Service.

5 Encyclopedia	53 How to hyphenate the word <i>cinema</i>	53()
	54 Amendments to the Constitution	54()
	55 A discussion of standards of living	55()
	56 The population of a particular small town	56()

1 American history textbook	57 List of news dispatches on CCC activities	57()
2 Book of quotations	58 A short account of the early history of Manhattan Island	58()
3 Library catalog	59 The author of <i>Neither in the Street</i>	59()
4 <i>National Geographic Magazine</i>	60 Information about the growth of slavery in the United States	60()
5 <i>New York Times Index</i>	61 Who said, 'Brevity is the soul of wit'	61()
	62 Pictures and story of recent developments in the TVA	62()

1 Daily newspaper	63 The Pulitzer Prize awards of 1930	63()
2 <i>Readers' Guide to Periodical Literature</i>	64 Today's price quotations on stocks and bonds	64()
3 <i>Time</i>		
4 <i>World Almanac</i>		
5 Library catalog		

Cooperative French Test Junior Form⁶¹

DIRECTIONS Each of the English sentences and phrases below is followed by a translation in which there is a blank indicated in this way (____). The translation will be correct when one of the five numbered words, phrases or endings listed at the left of the group is inserted in the blank (____). Decide which of the five items will make the translation complete and correct and put its number in the parentheses at the right hand edge of the page.

IV

1 ce	29 These books	(____) livres	()
2 ces			
3 cet	30 That school	(____) école	()
4 celles			
5 cette	31 That money	(____) argent	()

VIII

1 qui	38 What are they ask		
2 quoi	ing for?	(____) demandent-ils?	()
3 quelles	39 Who came down the	(____) est descendu	
4 que	first?	le premier?	()
5 qu	40 Which roads are the	(____) routes sont les	
	best?	meilleures?	()

XIII

1 -se	50 They lighted several	On a allumé plusieurs feu	
2 -es	fires	(____)	()

⁶¹ Devised by Jacob Greenberg and Geraldine Spaulding and published by Cooperative Test Service 1936

3 -x	51 I didn't buy the	Je n'ai pas acheté les autr	
1 -s	other books	() livres	()
5 No	52 He had black hair	Il avait les cheveux ()	()
ending		noirs	()
needed			

Sones Harry High School Achievement Test²²

SECTION G [MATHEMATICS] IMPORTANT THEOREMS IN GEOMETRY

DIRECTIONS In the parentheses after each geometric condition given below in Column 2 write the number of the results in Column 1 that could be proved by it

COLUMN 1 (RESULTS)	COLUMN 2 (CONDITIONS)
1 angles equal	66 If two opposite sides are equal and parallel () 66
2 triangles congruent	67 If perpendicular to the same line () 67
3 triangles similar	68 If the sides are proportional () 68
4 lines perpendicular	69 If they have equal arcs () 69
5 lines parallel	70 If side-angle-side equal side-angle-side respectively () 70
6 quadrilateral is a parallelogram	71 If they are parallelograms with equal bases and altitudes () 71
7 parallelogram is a rectangle	72 If their central angles are equal () 72
8 two arcs equal (in same or equal circles)	73 If a tangent is drawn to the radius at point of contact () 73
9 two chords equal (in same or equal circles)	74 If corresponding parts of congruent triangles () 74
10 areas of polygons equivalent	75 If one angle is a right angle () 75

II Rules and Suggestions for Construction

The purpose of the first three suggestions is to avoid irrelevant clues and that of the remaining five is to reduce the amount of time required to take the test

1 Include only homogeneous material in each matching exercise Do not mix in a single test such dissimilar associations as persons and events dates and events terms and definitions Put short titles at the top of both columns to describe the contents accurately For example Column 1 *Events* Column 2 *Dates*

2 Check each exercise carefully for unwarranted clues that may indicate matching pairs For each item ask yourself this question What is the least amount of information that must be known in order to select the right response?

3 Avoid making the test too easy The difficulty of a matching exercise may be increased by including more responses than needed and by using some of the responses more than once in the same test

4 One list should consist of single words numbers or brief phrases In general the column of short terms should contain the items from which the choice is made

5 The items in the response column should be arranged in systematic order If the list consists of dates they should be in chronological order For other items alphabetical order will assist the pupils in locating the desired response The responses in the column should then be numbered consecutively

²² Devised by W. W. D. Sones and David P. Harry, Jr. and published by World Book Company

6 *Indicate clearly the basis upon which matching is to be done* This should be specified both in the directions and in the column headings The pupil will be told to put the NUMBER of the response selected in the answer space beside the test item

7 *The matching exercise should contain at least five and not more than fifteen items* Larger lists waste time and shorter lists increase the possibility of guessing the correct response

8 *All the items for the matching exercise should be on a single page* Turning the page back and forth in search of desired responses is both confusing and time-consuming

G. Rearrangement Tests

Appendix C, pages 454-455 contains remarks about the preparation and scoring of rearrangement (ranking, sequence, chronology, continuity) items and should probably be consulted at this point Table 48 on page 455 makes scoring them rather simple, thereby eliminating one of the chief objections to this item type

Part II of the Allport-Vernon-Lindzey "Study of Values"⁶³ consists of 15 ranking items, the examinee being asked to "Arrange these answers in the order of your personal preference by writing, in the appropriate box at the right, a score of 4, 3, 2, or 1 To the statement you prefer most give 4, to the statement that is second most attractive 3, and so on " Two of these items are

2 In your opinion can a man who works in business all the week best spend Sunday in—

a trying to educate himself by reading serious books

a

b trying to win at golf, or racing

b

c going to an orchestral concert

c

d hearing a really good sermon

d

13 To what extent do the following famous persons interest you—

a Florence Nightingale

a

b Napoleon

b

⁶³Devised by Gordon W. Allport, Phillip E. Vernon, and Gardner Lindzey and published by Houghton Mifflin Company, 1921.

c Henry Ford



d Galileo



SELECTED REFERENCES FOR FURTHER READING

- Adkins, Dorothy C, and others, *Construction and Analysis of Achievement Tests* Washington, D C U S Government Printing Office, 1947 292 pages
- Buros, Oscar K (Editor), *The Fourth Mental Measurements Yearbook* Highland Park, New Jersey Gryphon Press, 1953 1163 pages
- Elbel, Robert L, "Writing the Test Item," Chapter 7 in E F Lundquist (Editor), *Educational Measurement* Washington, D C American Council on Education, 1951.
- Fzell, L B, "A Device for Scoring Chronology Tests," *Social Education*, 13 329-331, November, 1949
- Goodenough, Florence L, *Mental Testing* New York Rinehart & Company, 1949 Chapter 9, "The Analysis and Selection of Test Items"
- Henry, Nelson B (Editor), "The Measurement of Understanding," *Forty Fifth Yearbook of the National Society for the Study of Education, Part I* Chicago University of Chicago Press, 1946 338 pages
- Jordan, A M, *Measurement in Education* New York McGraw Hill Book Company, 1953 Chapter 3, "Constructing Achievement Tests"
- Odell, C W, *How to Improve Classroom Testing* Dubuque, Iowa Wm C Brown Company, 1953 156 pages
- Stephenson, William, *Testing School Children* New York Longmans, Green and Company, 1949 Chapter VI, "Tests of Creative Imagination," and Chapter VII, "Performance Tests"
- Travers, Robert M W, *How to Make Achievement Tests* New York Odyssey Press, 1950 180 pages
- Traxler, Arthur E, Jacobs, Robert, Selover, Margaret, and Townsend Agatha, *Introduction to Testing and the Use of Test Results in Public Schools* New York Harper & Brothers, 1953 Chapter 4, "How Can Tests Be Selected?"
- Weitzman, Ellis, and McNamara, Walter J, *Constructing Classroom Examinations—a Guide for Teachers* Chicago Science Research Associates, 1949 153 pages

The Construction and Use of Essay Examinations

Stalnaker¹ compares the merits of essay and objective tests in a thorough and impartial manner, concluding that both have considerable value when properly used. The summary to his chapter on page 530 is especially interesting.

The essay test has been the subject of repeated and often unfair attacks by psychologists and educationalists interested in the measurement of achievement as a science. As a result, the essay test remains largely undeveloped, although it continues to be used widely by the classroom teacher. The values claimed for it have not been generally established, yet it may well be a basic test form which properly controlled can measure important outcomes of learning not yet otherwise measured. It also has other potential values which have been described. It has several important and unique advantages as an educative influence. The fact that it continues to be a test form widely used by the teacher preparing his own test would alone seem to justify further development and research.

For many years the College Entrance Examination Board² has been concerned with the improvement of essay tests, particularly the increasing of scoring agreement on English compositions. Its journal, *The College Board Review*, published three times yearly, and the *Annual Report of the Director*³ contain valuable reports of work with essay tests.

To limit the use of informal teacher-made tests to those classified as objective in type is an unwarranted restriction. The so-called traditional

¹ John M. Stalnaker, "The Essay Type of Examination," Chapter 13 in E. F. Lindquist (Editor), *Educational Measurement*. Washington, D. C.: American Council on Education, 1951.

² Abbreviated CEEB and located at 425 West 117th Street, New York 27, N. Y.

³ The 53rd annual report covers the period October 1, 1952-September 30, 1953.

test or essay examination still has a legitimate place in the modern school. This chapter will consider some of the advantages and limitations of this type of test, and offer suggestions for its improvement and use.

A. Limitations of the Essay Examination

As ordinarily employed, the essay examination has certain serious limitations. It suffers in comparison with most forms of objective tests on the three important criteria of a satisfactory measuring instrument, validity, reliability, and usability.

Low validity. In the first place, the essay examination as commonly used has low validity. Several factors contribute to this condition. The limited sampling of the essay examination is often pointed out. Ruch,⁴ for example, produced evidence to show that the essay called forth less than half the knowledge the average pupil actually possessed on the subject as determined by objective tests, and required twice the time to do it. The essay also includes many irrelevant factors, such as the quality of the spelling, handwriting and English used, as well as bluffing, for which no correction formulas exist. It has been suggested that the essay overrates the importance of knowing how to say a thing and underrates the importance of having something to say. In view of these limitations, the ordinary essay examination has little validity as an instrument of diagnosis.

Low reliability. In the second place, the essay examination as commonly used is low in reliability. Since short tests are usually less reliable than long tests, the narrow sampling afforded by essay examinations would tend to restrict its reliability. Still more serious is the subjectivity of scoring. Numerous studies have shown that teachers cannot agree with each other as to the values to be allowed examination papers of the essay type. Studies have also shown that the same teachers cannot agree with themselves on a second series of values assigned independently to the same papers. Part of this is due to different standards of marking and different weighting of the questions. Certain other factors such as the physical and mental condition of the person marking the papers also tend to condition the mark assigned a paper by a given teacher at any particular time. An English poet states the situation as follows:⁵

"Twixt Right and Wrong the Difference is dim
Tis settled by the Moderator's Whim
Perchance the Delta on your Paper marked
Means that his Inch has disagreed with him

In a study⁶ made at the University of West Virginia, Ashburn came to

⁴G. M. Ruch, *The Objective or New Type Examination*, page 54. Chicago: Scott Foresman & Company, 1929.

⁵Quoted by I. L. Kandel, *Examinations and Their Substitutes in the United States*, page 28. New York: Carnegie Foundation for the Advancement of Teaching, 1936.

⁶Robert R. Ashburn, "An Experiment in the Essay Type Question," *Journal of Experimental Education*, 7:13, September, 1938.

the conclusion that "the passing or failing of about 40 per cent depends, not on what they know or do not know, but on *who* reads the papers" and that "the passing or failing of about 10 per cent depends on *when* the papers are read" It has been observed that the scores tend to rise as time passes, and that the values assigned tend to be greatly influenced by those allowed the paper immediately preceding For example, one writer asserts that, "A *C* paper may be graded *B* if it is read after an illiterate theme, but if it follows an *4+* paper if such can be found, it seems to be of *D* caliber"⁷

That this situation is not peculiar to American education is indicated by the Examination Inquiry conducted by the International Institute of Teachers College, Columbia University⁸ In fact, one writer⁹ asserts that evidence showed the unreliability of essay examinations in Europe was "even more serious" than had been revealed many times in America In support of this rather surprising conclusion, he says "In the English studies, examiners were found to reverse their judgments almost completely when asked to mark the same papers they had scored a year before"

Bowles¹⁰ comments concerning England's examination system for college entrance are illuminating He concludes that it is quite deficient when judged by American standards of reliability and statistical validity, but that because of various safeguards for the individual "the system works" well

In fairness to the essay examination, however, it should be pointed out that many of the studies reported have been with unimproved forms of the examination given under unfavorable conditions Often the essay examination at its worst has been compared with an improved objective test Under such conditions, the former is bound to show up in an unfavorable light If objective tests had been scored under similar conditions, without scoring rules or keys, the agreement of the scores would be less impressive As a matter of fact, even with scoring rules and keys, the agreement among the scores on objective tests allowed by amateur scorers is far from perfect Under favorable conditions the agreement among scorers of essay examinations approximates that reported for objective tests One study¹¹ reports that the average correlation coefficient between first and second scorings of an essay test in history by three experienced scorers was .98 Another

⁷ John M. Stalnaker 'The Problem of the English Examination' *Educational Record* 17 41 Supplement No. 10 October 1936

⁸ Published by the Bureau of Publications 1936

⁹ W. Carson Ryan Jr. The Seventh World Conference of the New Education Fellowship II *School and Society* 44 364 September 19 1936

¹⁰ Frank H. Bowles *The College Entrance Examination Board 51st Annual Report of the Director 19 1* pages 23-30 College Entrance Examination Board 425 West 117th Street New York 27 New York 1932

¹¹ Roy E. Cochran and Charles C. Weidemann Improvement of Consistency of Scoring the Explain and Discuss Essay Examination a paper read before Section C of the American Educational Research Association at Cleveland Ohio March 1 1939

study¹² reports that the median coefficient obtained between two independent readings of certain College Entrance Board examinations was .97. All twenty of the coefficients were above .90, with the exception of English, which was .81. It must be kept in mind, however, that these examinations were so worded as to make the scoring more objective than is usually possible with ordinary essay examinations.

It should be noted that most studies having to do with the reliability of essay examinations really show the reliability of *marking the examination* rather than the reliability of the examination itself. A few studies have been reported of the correlation between two forms of an essay examination designed for a particular purpose which were given to the same pupils and carefully marked by experienced examiners. McGregor and Ruch¹³ used this procedure in studying eighth grade examinations in sixteen subjects from 952 pupils in eleven states. Each paper in the two sets of examinations was marked independently by two experienced teachers. This study made it possible to compare the reliability of the examination with the reliability of *marking the examination*. The agreement of the two independent markings of the same papers is represented by an average correlation of .62, while the agreement of the two sets of examinations marked by the same teacher is represented by an average correlation of only .43. One of Ruch's students, Dr. W. E. Gordon¹⁴ made a similar study of the New York Regents' Examinations with startlingly comparable results. He found the average agreement of the two independent markings of the same papers was .72, while the average agreement of the two sets of examinations marked by the same teacher was only .42. Another study¹⁵ conducted at the University of Chicago High School showed that two independent sets of marks assigned by two experienced readers of essay examinations agreed to the extent of .944 on Form A and .845 on Form B, but that the correlation between Form A and Form B was only .60. These three studies seem to agree on one important point: *The reliability of marking the essay examination is higher than the reliability of the examination itself.*

Low usability. The essay examination also ranks low in usability. There seems no escape from the fact that this type of examination is time consuming, both for the pupil and for the teacher. In fact, the additional expenditure of time and energy over that needed for objective tests is so serious a limitation that the use of essay examinations can be justified only if it can be shown that the values realized are commensurate with this investment.

¹² John M. Stalnaker, *Essay Examinations Reliably Read*, *School and Society* 46: 671-672, November 20, 1937.

¹³ G. M. Ruch, *The Objective or New-Type Examination*, pages 91-96, Chicago: Scott Foresman & Company, 1929.

¹⁴ *Ibid.*, pages 97-98.

¹⁵ Arthur E. Traxler and Harold A. Anderson, *The Reliability of an Essay Examination*, *Journal of Educational Psychology* 43: 534-535, September 1935.

B. Advantages of the Essay Examination

Reliability and usability. Even the most enthusiastic advocate of essay examinations would scarcely claim their superiority over objective tests on the grounds of reliability or usability. The best that can be hoped for essay examinations is that by the use of improved techniques their reliability may approach that of objective tests. As regards usability, the fact that the questions can be written on the blackboard is an advantage only in those schools which lack duplicating facilities. The reduction in time required to prepare essay examinations is more apparent than real, if the work is well done. Whatever advantage arises therefrom is more than offset by such considerations as the extra time demanded for giving and scoring.

Validity. It is apparent that if the use of essay examinations can be justified it must be upon the ground of their superior validity for certain purposes. What, then, are the unique functions of these examinations?

Unfortunately, upon this crucial issue little experimental evidence exists. One study¹⁶ indicated that about 30 to 40 per cent of the mental functions measured by improved essay tests of the "compare and contrast" type were not measured by true-false tests covering the same material. Two similar studies by Cochran and Weidemann¹⁷ compared one-word fact tests and essay tests of the improved "explain" and "discuss" types covering the same material, and concluded that about 40 per cent of the mental functions measured by the latter were not measured by the former. The important question of just what unique mental functions each type of test measures remains to be answered.

In the absence of experimental evidence, it is necessary to fall back on logical considerations. The essay test appears to be useful for measuring four objectives of instruction: functional information, certain aspects of thinking, study skills and work habits, and a functioning social philosophy. It will be noted that these objectives emphasize the *functioning* rather than the mere possession, of knowledge.

There would appear to be little justification for using essay tests for the recall of knowledge in piecemeal fashion. Sims,¹⁸ however, analyzed 458 questions ordinarily classified as of the essay type, and found that fewer than half in the high school and fewer than one in five in the elementary school involved discussion, the others being almost equally divided between simple-recall and short answer questions requiring not more than

¹⁶ C. C. Weidemann and Lyndall Fisher Newens, "Does the 'Compare-and-Contrast' Essay Test Measure the Same Mental Functions as the True-False Test?" *Journal of General Psychology*, 9, 430-449, October, 1933.

¹⁷ Roy E. Cochran and Charles C. Weidemann, "'Explain' Essay vs. Word Answer Fact Test," *Phi Delta Kappan* 17, 59-61, 75, December, 1934, and "A Study of Special Types of Tests," *Phi Delta Kappan* 19, 113-115, 131, January, 1937.

¹⁸ Verner Martin Sims, "Essay Examination Questions Classified on the Basis of Objectivity," *School and Society* 35, 100, 02, January 16, 1932.

one sentence for a response. The Evaluation Committee of "the Seattle Schools" came to the conclusion that the evaluation of growth in language ability would require the use of several types of tests.

Both objective and essay tests appeared necessary to measure achievement at the various levels of knowledge. Objective tests of the multiple-choice and matching type could be used to measure achievement at the *recognition and recall* levels. However, evaluating achievement at the level of *interpretation and evaluation* would require essay-type tests, as well as certain kinds of objective tests. Evaluating achievement at the level of *application* would seem to be done most effectively by essay tests, since this would involve measuring the student's ability to utilize information learned in one situation in the solution of problems in a new setting.

One other advantage of the essay examination should be mentioned. Several experimental studies have shown that the type of measurement used by the teacher influences the type of study procedures employed by the pupils.²¹ When pupils expect the test to be of the essay type, in whole or in part, they seem more likely to employ such desirable study techniques as making outlines and summaries, and seeking to perceive relationships and trends, than is done when objective tests are used exclusively.

The practical conclusion follows that neither the essay nor objective test should be used exclusively. From Lee and Segel's²² analysis of the measurement practices of 1,600 secondary school teachers, distributed widely over the United States, and from the Hensley-Davis study²³ it appears that teachers favor the use of a combination of the two types. It is encouraging that the practice of more and more teachers seems to be governed by the sound philosophy of measurement stated by Lindquist in the following sentence:²⁴

The intelligent point of view is that which recognizes that whatever advantages either type may have are specific advantages in specific situations that while certain purposes may be best served by one type other purposes are best served by the other, and above all that the adequacy of either type in any specific situation is much more dependent upon the ingenuity and intelligence with which the test is used than upon any inherent characteristic or limitation of the type employed.

C. Suggestions for Improving Essay Examinations

Although the essay examination has been in existence for hundreds of years, the amount of research devoted to it is much less than that devoted

²¹ Helen F. Olson. Evaluating Growth in Language Ability. *Journal of Educational Research* 39: 247. December 1945.

²² For a fuller discussion of this point see Chapter 11.

²³ J. Murray Lee and David Segel. *Testing Practices of High-School Teachers*. page 38. United States Office of Education Bulletin No. 9. 1936.

²⁴ Iven H. Hensley and Robert A. Davis. What High School Teachers Think and Do About Their Examinations. *Educational Administration and Supervision* 38: 219. 228. April 1952.

²⁵ Herbert E. Hawkes, F. F. Lindquist and C. R. Mann. *The Construction and Use of Achievement Examinations*. page 20. Boston: Houghton Mifflin Company. 1936.

to the objective test which is comparatively new. Furthermore, practically all the research relating to the former has been of a negative kind. Its purpose has been to show how poor unimproved essay examinations are, rather than to devise means for their betterment. However, a study of the meager experimental literature does yield several positive suggestions. The next two sections will be devoted to a consideration of some of the most promising of these suggestions.

Improving the construction and use of essay examinations. It is just as important to know *where* to use the essay examination as it is to know *how* to use it. It is wise to restrict the use of the essay test to the measurement of those functions for which it is best adapted. There would usually appear to be no good reason for employing subjective measurement where objective measurement is adequate. What, then, does the essay examination attempt to do?

Weidemann²⁴ recognizes eleven definable types of improved essay examinations. Arranged in a series from simple to complex, these types are as follows: (1) *what*, (2) *who*, (3) *when*, (4) *which*, and (5) *where*, (6) *list*, (7) *outline*, (8) *describe*, (9) *contrast*, (10) *compare*, (11) *explain*, (12) *discuss*, (13) *develop*, (14) *summarize*, and (15) *evaluate*. The first two types seem hardly distinguishable from recall tests of the objective type. Many years ago Monroe and Carter²⁵ made a very suggestive classification of thought questions into twenty types. These types, together with an illustration of each, taken from the field of measurement, appear below.

Thought Questions

- 1 Selective recall—basis given
Name three important developments in measurement which occurred during the first decade of the twentieth century.
- 2 Evaluating recall—basis given
Name the three persons who have had the greatest influence on the development of intelligence testing.
- 3 Comparison of two things—on a single designated basis
Compare essay examinations and objective tests from the standpoint of their effect upon the study procedures used by the learner.
- 4 Comparison of two things—in general
Compare standardized and non standardized tests.
- 5 Decision—for or against
In which, in your opinion, can you do better, oral or written examinations? Why?

²⁴ C. C. Weidemann, "Written Examination Procedures," *The Delta Kappa* 16: 78-83, October 1933; also C. C. Weidemann, "Review of Essay Test Studies," *Journal of Higher Education* 12: 41-44, January 1941.

²⁵ Walter S. Monroe and Ralph E. Carter, *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students*, 26 pages, Urbana, Illinois: Bureau of Educational Research Bulletin Number 14, University of Illinois, 1923.

- 6 Cause or effects
How do you account for the great popularity of objective tests during the last thirty-five years?
- 7 Explanation of the use or exact meaning of some phrase or statement in a passage
What is the meaning of "Delta" in the verse quoted on page 193?
- 8 Summary of some unit of the text or of some article read
Summarize in not more than one page the advantages and limitations of essay examinations
- 9 Analysis (The word itself is seldom involved in the question)
Why are many so-called "progressive educators" suspicious of standardized tests?
- 10 Statement of relationships
Why is it that nearly all essay examinations, regardless of the school subject, tend to be measures of the learner's mastery of English?
- 11 Illustrations or examples (your own) of principles in science, construction in language, etc
Give two original examples of specific determiners in objective tests
- 12 Classification (usually the converse of No. 11)
What type of error appears in this test item? "With what Balkan country did the Allies fight in World War I?"
- 13 Application of rules or principles to new situations
In the light of China's experience with state examinations what would you expect to be the effect of the Regents' Examinations in New York?
- 14 Discussion
Discuss the place of measurement in science
- 15 Statement of aim—author's purpose in his selection or organization of material
In view of the author's discussion on pages 19 and 20, why are so many authorities quoted in Chapter 1?
- 16 Criticism—as to the adequacy, correctness, or relevancy of a printed statement, or a classmate's answer to a question on the lesson
Criticize or defend the statement "The essay examination overrates the importance of knowing how to say a thing and underrates the importance of having something to say."
17. Outline
Outline the principal steps in the construction of an informal teacher-made test
- 18 Reorganization of facts (a good type of review question to give training in organization)
Name ten practical suggestions from Chapters 4, 5, and 6 that are particularly applicable to the subject you teach or plan to teach
- 19 Formulation of new questions—problems and questions raised
What are some problems relating to the use of essay examinations that require further study?
- 20 New methods of procedure
Suggest a plan for proving the truth or falsity of the contention that exemption from examinations is a good policy in high school

Special advantages. It will be noted that the classifications by Weidemann and by Monroe and Carter recognize a considerable number of rather distinct abilities, which are measurable by essay tests. It is probably best to measure each one separately rather than to attempt to measure several of them by the same test. It will be further noted that the emphasis in most of these types is upon organization, relationship, evaluation, application or some similar ability to which a purely objective test may be poorly adapted. Teachers should study each type of essay question carefully until they are familiar with its distinguishing characteristics. If a proposed essay question does not seem to conform to one of these types, it had usually better be reworded or adapted to some form of the objective test. No question should be included until its purpose has been clearly defined.

The essay examination would appear to be particularly valuable in two situations. The first of these is obviously in such courses as English composition and journalism, where the student's ability to express himself effectively is the major objective of instruction. The second situation is in advanced courses of other subjects, where critical evaluation and the ability to assimilate and organize large amounts of material constitute important objectives. In this connection it is significant to note that Jones²⁶ found that 68 per cent of the college students who took senior comprehensive examinations and 55 per cent of the superior students in other colleges stated their views as follows: "I think one's ability is far better shown through discussion questions than through short objective questions."

There is some evidence that a more valid sampling of the pupil's knowledge is afforded by increasing the number of questions and reducing the length of discussion expected on each. In many cases a well-constructed paragraph is sufficient. Very few discussions need exceed one or two pages in length. In any case, the question should be so worded as to restrict the responses toward the objective which it is desired to measure. For example, Wrightstone suggests that the question, "Explain the reasons for the strike at General Motors in 1937," is too general, and would be improved if it were restricted by the addition of the phrases "to show (a) the labor grievances of the employees, (b) the practices of the employer (c) related national, social, and economic factors, (d) the rival labor unions and (e) the method of striking."²⁷ It must be recognized, however, that such suggestions at least in part, take from the essay examination its uniqueness. The proposed modifications may appear to improve the reliability of the traditional examination by the obvious device of making it more like the objective test.

²⁶ Edward Safford Jones, *Comprehensive Examinations in American Colleges*, page 373. New York: The Macmillan Company, 1933.

²⁷ J. W. Wrightstone, "Are Essay Examinations Obsolete?" *Social Education* 1: 403, September 1937.

One of the difficulties with constructing essay tests is that the process appears so easy. As a matter of fact it is *probably more difficult to construct essay tests of high quality than it is to construct objective tests of high quality*. Much care and thought must be given to their construction if tests of any kind are to measure anything but mere memory for factual knowledge. Many of the general principles of testing outlined in an earlier chapter are applicable to essay tests as to objective tests. There is always risk that in attempting to phrase essay questions so that they can be scored more objectively, the result may be made less satisfactory than an out-and-out objective test. Critical revision utilizing if possible the judgment of a colleague is especially important.

Preparing students to take essay tests. Some writers have emphasized the importance of training pupils in taking examinations. Worcester²⁸ suggests that the essay examination is 'obviously invalid and unfair at least in part because the pupils are being required to take a test on a type of work for which they have had no specific training. The rational solution offered is to supply the necessary training rather than to abandon the essay examination. Wider experience and training in preparing for and in taking tests of all types is likely to increase the accuracy of measurement. Edmiston²⁹ prepared instructions to pupils for taking examinations which were far more elaborate than the usual directions accompanying tests. He found that the use of these instructions increased the validity of the examinations and produced definite improvements in students' records of achievement from examinations.' It would appear wise to provide instruction of this sort in the regular program of studies. It is most unfortunate when a pupil fails to receive recognition for knowledge he actually possesses simply because he has not mastered the technique of putting it on paper. Edmiston's suggestions³⁰ given below will prove helpful in planning such a program of instruction.

IMPORTANT CONSIDERATIONS IN TAKING EXAMINATIONS

- 1 Your name should appear on the first or last sheet of the examination if sheets are securely bound. Each loose sheet should have the name entered inconspicuously, preferably on back where it will not be seen by the scorer when scoring.
- 2 Write legibly. Your answer can't be right if it can't be read. Be sure your pen or pencil (if allowed) fosters distinct and not blurred writing.
- 3 Use terms or a vocabulary suited to the subject. Do not use a word unless its meaning is clear to you and repeat a word rather than use another which may not have exactly the same desired meaning.

²⁸ D. A. Worcester, On the Validity of Testing, *School Review* 42: 5-7, 531, September 1934.

²⁹ R. W. Edmiston, Examine the Examination, *Journal of Educational Psychology* 30: 126-138, February 1939.

³⁰ *Ibid.* pages 137-138.

4. Space (the back of sheets, the margins, or an extra sheet) should be used for
 - a. computations.
 - b. practice in the formation of desirable statements, not padded but furnishing quality rather than quantity to the answer.
 - c. the hasty jotting of facts pertaining to some questions when these facts arise, while working upon another question.
5. The statement of each question must be fully considered. Carelessness not only penalizes the student but also lowers the dependability of the measurement obtained by the instructor.
6. The directions telling how to answer the questions should be carefully followed. Underscore the important points in the directions.
7. In essay questions, *underscore* the part of the statement that furnishes the direct question asked. Then underscore any parts of the statement which furnish data for the answer. Number each part so that you will not omit anything from your answer.
8. Proceed directly through the examination with no lengthy consideration of unfamiliar points. After completing the parts which were readily answered, start again and answer those questions which yield to more diligent effort. Do not waste time by trial and error method upon questions which bring no recognition or recall of related materials. After completing the second consideration of the test, spend the remainder of the time upon the more familiar of the unanswered questions. Note that hesitation wastes time, ruins confidence, and destroys mind-set.
9. If after thorough consideration you do not understand some direction or question due to other than lack of knowledge of the course, call the attention of the person in charge with as little disturbance as possible in order that the tester may come to your seat or allow you to come to him as conditions may determine.
10. Reread each answer before passing to the next question and the completed examination before delivery to the instructor. Is the meaning clear and writing legible?

By way of summary, three important suggestions for the construction and use of essay examinations are as follows:

1. Restrict the use of the essay examination to those functions to which it is best adapted. When it is not clear that the essay type is required for measuring the desired objective, use the objective test.
2. Increase the number of questions asked and reduce the amount of discussion required on each. Always indicate clearly the type of discussion desired.
3. Make definite provisions for teaching pupils how to take examinations. Specific training in preparing for and in taking tests and examinations of the various types commonly encountered is a legitimate objective of instruction.

Improving the scoring or grading of essay examinations. Rinsland²¹ makes a distinction between the terms *scoring* and *grading*. Scoring

²¹ Henry Daniel Rinsland, *Constructing Tests and Grading in Elementary and High School Subjects*, page 302. New York: Prentice-Hall, Inc., 1938.

is an objective process of counting right or wrong responses, whereas grading always means interpreting quality in terms of some criterion. Strictly speaking, then, it is more correct to speak of grading or rating essay examinations than it is to speak of scoring them.

It is, of course, apparent that whatever claims are made for the validity of the essay test as a measuring instrument are conditioned upon the assumption that the papers can be read accurately. Not only must the essay test, for example, call forth from superior pupils responses which are consistently superior, but the teachers marking the papers must be able consistently to recognize that they are superior responses. The same is true of responses with other degrees of merit. The grading of the essay examination, therefore, occupies a strategic position.

To begin with, certain preventive measures are important. A careful wording of the questions and directions to the pupil which indicate clearly just what type of response is expected will simplify the problem of marking the papers. *The use of optional questions should be discouraged*²². The simple precaution of having the pupil record his name inconspicuously either on the back or at the end of the paper, rather than at the top of each page, is likely to increase the accuracy with which the paper is graded.

Cochran and Weidemann²³ outline a procedure for evaluating essay examinations, the essentials of which can be taught in ten minutes. This is shown by the fact that the majority of the consistency coefficients of two series of scorings made five weeks apart were between .80 and .90 for teachers with ten minutes of training. Independent scores by experienced readers showed an average agreement of .98 when the procedure given below in a slightly modified and abridged form was used.

SUGGESTIONS FOR MARKING ESSAY EXAMINATIONS (After Cochran and Weidemann)

- 1 I read over a sampling of the papers to obtain a general idea of the grade of answer I may expect.
- 2 I score one question through all of the papers before I consider another question. I have found two outstanding advantages in scoring one question through an entire set of papers. The first is that the comparison of answers appears to make the grades more exact and just. The second is that having to keep only one list of points in mind saves time and promotes accuracy.
- 3 Before scoring any papers I read the material in the text which covers the questions and also the lecture notes on the subject.
- 4 I make a list of the main points which should be discussed in every answer. Each of these points must be weighed and assigned a certain value if the scoring is to approach accuracy. This value assigned to the main points needed for a reasonably adequate answer is designated as the minimum score. If a pupil elaborates and discusses points not required yet pertinent to the question his answer is given an additional value, called the extra score. This extra score may vary for different pupils but may not exceed a certain set maximum.

²² John M. Stalnaker, 'The Essay Type of Examination' *op cit* pages 505-506

²³ Roy I. Cochran and C. C. Weidemann *op cit*

5. After the points have been weighed, the actual scoring begins. I read the answer through once and then check back over it for fact details. I attempt to mark every historical mistake on the paper and write in briefly the correction. As I read the answer I make a mental note of the points omitted and the value of each point, so that when the end of the question is reached, I have the minimum grade figured. If there is any additional or extra percentage to be given, it is added to the minimum score, and then the value of the question is written in terms of the per cent deducted rather than the positive per cent. Then when every question on a paper is scored, it is a simple matter to add the negative quantities and obtain the final grade.

It is difficult to overemphasize the importance of three things: (1) the preparing in advance of a list of answers which are considered adequate for the objectives of the test; (2) the assigning of a specific value to each essential part of the answers; and (3) the grading of one question through all the papers before going on to another question. Most students of the problem recommended attempting to distinguish a relatively small number of degrees of merit in an answer. Perhaps as good a plan as any is to allow credit for each part of the answer considered essential to a question as follows: 3 for superior, 2 for average, 1 for inferior, and 0 for an omission or wrong reply. Stalnaker²⁴ found that the weighting of essay questions was of negligible value—the correlation between weighted and unweighted scores on the College Entrance Board Examinations varying from .97 to .997.

Grading by sorting. In addition to the points made by Cochran and Weidemann, several authorities have found another suggestion helpful. The suggestion is to make a sorting of the papers into three to five piles, according to the merit of the discussion of each question on the basis of a brief preliminary examination of the answers. Sims describes one such procedure as follows:²⁵

1. Quickly read through the papers and on the basis of your opinion of their worth sort them into five groups as follows: (a) very superior papers, (b) superior papers, (c) average papers, (d) inferior papers, (e) very inferior papers.
2. Reread the papers in each group and shift any that you feel have been misplaced.

Flanagan²⁶ has shown that the optimal percentages for five groups are 9, 20, 42, 20, and 9. Therefore, about 10 per cent of the papers might be called "very superior" and 10 per cent "very inferior." Twenty per cent

²⁴ John M. Stalnaker, "Weighting Questions in the Essay-Type Examination," *Journal of Educational Psychology*, 29: 481-490, October, 1938.

²⁵ Verner Martin Sims, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," *Journal of Educational Research*, 24: 216-223, October, 1931.

²⁶ John C. Flanagan, "The Effectiveness of Short Methods for Calculating Correlation Coefficients," *Psychological Bulletin*, 49: 342-348, July, 1952.

would be "superior" and a like percentage would be "inferior." The remaining 10 per cent are "average." These are rough approximations, of course, dependent upon the ability level of the particular student group being graded.

The preliminary sorting of the papers into piles of approximately equal merit before assigning numerical values to them will help to avoid the difficulty pointed out by Stralaker namely, that the values allowed a paper are often greatly influenced by the merit of the paper which happens immediately to precede it in the order of scoring. It is also easier to locate papers distinctly out of line with those in a particular group supposedly of similar quality. It is a good idea to throw the papers into a single group after each question has been evaluated and before they are re-sorted into piles according to the merits of the discussions of the next question. This procedure will make it easier to conceal the identity of the particular pupil whose paper is being judged and so to avoid one of the most disturbing factors in marking essay examinations.

The school should adopt a policy regarding what factors shall be considered, and what factors shall not be considered, in evaluating a written examination. *Only those factors should be taken into account which afford evidence of the degree to which the pupil has attained the objectives set up for that particular course.* Except in English classes this will rule out making arbitrary reductions for such things as faulty sentence structure, paragraphing, handwriting and the spelling of nontechnical words. These factors will be considered only in so far as they affect the clarity of the pupil's discussion. It is always legitimate to hold the pupil responsible for the spelling, as well as the meaning of the vocabulary which is peculiar to the course.

This does not mean that the quality of the written English used in examinations is unimportant and should therefore be disregarded. On the contrary, it is always very important. But it should be considered only in relation to that for which it may be accepted as valid evidence namely, in determination of the pupil's mark in English. Where the teacher has complete charge of an entire grade, this adjustment is easy to make. But where the school is departmentalized the problem is more difficult. Even here it should be possible to work out a system whereby at intervals the papers in other subjects, after having been graded as to content, may be turned over to the English teacher to be judged from the viewpoint of their merits as English compositions. In this way it may be possible to sample the pupil's characteristic performance in written English better than when he writes a paper specifically for the English teacher. And what is equally important it makes the pupil's mark in other subjects a measure of achievement in those subjects rather than partly a measure of skill in English composition.

SELECTED REFERENCES FOR FURTHER READING

- College Board Review*, "Reading Conference," No 19 324-326, February, 1953
- Cook, Walter W, "The Functions of Measurement in the Facilitation of Learning," Chapter 1 in E F Lindquist (Editor), *Educational Measurement* Washington, D C American Council on Education, 1951
- Ellis, Albert, "An Experiment in the Rating of Essay-Type Examination Questions by College Students," *Educational and Psychological Measurement*, 10 707-711, Winter, 1950
- Flanagan, John C, "The Use of Comprehensive Rationales in Test Development," *Educational and Psychological Measurement*, 11 151-155, Spring 1951
- Henry, Nelson B (Editor), "The Measurement of Understanding," *Forty-Fifth Yearbook of the National Society for the Study of Education, Part I* Chicago University of Chicago Press, 1946 338 pages
- Kostick, Max M, and Nixon, Belle M, "How to Improve Oral Questioning," *Peabody Journal of Education*, 30 209-217, January, 1953
- Odell, C W, *How to Improve Classroom Testing* Dubuque, Iowa Wm C Brown Company, 1953 Chapters V and VI, "Discussion of Essay Examinations" and "Short-Answer Tests General"
- Remmers, H H, and Gage, N L, *Educational Measurement and Evaluation* New York Harper & Brothers, 1943 Chapter XII, "Essay Testing"
- Stalnaker, John M, "The Essay Type of Examination," Chapter 13 in E F Lindquist *Educational Measurement* Washington, D C American Council on Education, 1951
- Torgerson, Warren S, and Green, Bert F, Jr, "The Factor Analysis of Subject-Matter Experts," *Journal of Educational Psychology*, 43 354-363, October, 1952

PART III

The Testing Program

8

Steps in the Testing Program

Dear Professor

I have decided to give some tests in my school this fall. Please suggest a few good ones I might try. Also let me know where to get them and what they will cost.

Dear Professor

We gave the Up-to-Date General Achievement Tests at the beginning of the school year. As we now have most of them scored, please advise me how to use the results so as to get the most good out of them. Any help will be greatly appreciated.

Probably every college professor who offers courses in measurement has received letters like those above. They indicate that some school is undertaking, or has already undertaken, to use standard tests without understanding what it is all about. Always, testing should have a *program* to guide it.¹ What, then, is a "testing program"?

General considerations. The word "program" has certain important implications, such as *order, system, planning*. It implies a sequence of events that has been determined upon after careful thought, rather than some haphazard, hit-or-miss affair. One of the chief weaknesses of many attempts to use standard tests is that there has been no program worthy of the name. The whole procedure has simply led a precarious hand-to-mouth existence from beginning to end.

Spence² has suggested that "a good testing program should be supplementary *not* duplicative, usable *not* confusing, economical *not* burden

¹ Julian C. Stanley, "Standardized Tests and Educational Objectives," *Peabody Journal of Education* 28: 218-221, January 1951.

² Ralph B. Spence, "A Comprehensive Testing Program for Elementary Schools," *Teachers College Record* 34: 279-284, January 1933.

some, comprehensive *not* sporadic, suggestive *not* dogmatic, progressive *not* static " Such a program, at least in tentative form, may very well cover an extended period, rather than be adopted piecemeal year by year One advantage of this long-range planning is that it makes possible a varied program without leaving gaps or involving needless duplication Stenquist,² speaking from wide experience, strongly advocates "some sort of systematically recurring schedules as opposed to sporadic testing," since schedules make possible "enormously greater gains" from testing Spence offers for elementary schools what he calls "a conservative approach to the problem " This program is given in Table 26

TABLE 26

PLAN FOR A TESTING PROGRAM FOR THE ELEMENTARY SCHOOL (AFTER SIEGCE)

GRADE	INTELLIGENCE ^a	ACHIEVEMENT BATTERY ANNUAL (GIVEN IN MAR OR APR) ^b	ACHIEVEMENT TESTS FOR SPECIAL EMPHASIS ALL GRADES FROM 3 OR 4 TO 8 ROTATING (GIVEN IN OCTOBER) ^b
Kdg — I	Two Group Tests	Reading Battery	
II		Skill Subjects Battery	
III	One Group Test	Complete Battery	First Year—Reading
IV			Second Year—Arithmetic
V			Third Year—Social Studies
VI		Complete Battery	Fourth Year—Language
VII			Usage and Spelling
VIII		Complete Battery	Fifth Year—Reading, etc

^a Retests for special cases as needed preferably with an individual test

^b All dates based on groups beginning a grade in September Teachers use diagnostic tests throughout the year

It will be noted that this program calls for the use of both intelligence tests and achievement tests, and for the use of test batteries as well as of tests in the separate subjects It is also expected that the program will merely supplement rather than supplant the ordinary informal tests and examinations made by the classroom teacher A slight modification of the schedule as presented would involve giving a general test battery in all subjects about every third year, and an intensive program limited to one subject in each of the intervening years The cost of such a program of standard tests would be less than twenty-five cents per pupil per year If the tests are intelligently used, it is doubtful whether greater returns can be had by the school from the same amount of money spent in any other way

² John L Stenquist 'Recent Developments in the Uses of Tests,' *Review of Educational Research* 3 60 February, 1933

Traxler,⁴ in giving a practical discussion of the planning and administration of the testing program, divides tests into two broad categories. The first includes group tests of intelligence and achievement tests in the major subject matter areas. These should be administered at regular intervals to every normal pupil in the school. The second category includes individual intelligence tests, special aptitude tests, personality tests, and tests of vocational interest.

The following comprehensive 'Platform for the Use of Standard Tests' has been prepared by a committee of Massachusetts teachers.⁵

- 1 Scientific measuring instruments and the scientific method are badly needed in present educational practice. No business of the financial magnitude of education spends so little time and money for objective and scientific fact finding.
- 2 Standardized tests and measurements can fulfill their function of giving direction and efficiency to education only when used intelligently by teachers and administrators who have kept abreast of current knowledge on the subject and who are willing to follow the authors' directions for the administration and scoring of the tests used. The results of tests in which directions are not followed are worse than useless; they are misleading.
- 3 Every standardized test administered should be given for a specific purpose and having been given, its results should be used. Tests which are administered, scored, and piled in a cupboard serve no useful purpose.
- 4 Standardized tests can be used most efficiently when their use is planned over a long period of time.
- 5 Standardized tests have furnished valuable information to the school administrator in practically every instance in which they have been used. The possibilities of diagnostic tests in improving instruction through analysis and diagnosis of individual and class weaknesses have not nearly been realized. Tests are of the greatest value when their results cause a teacher to redefine his objectives, alter his methods, and redirect his emphasis as a result of new increased and more exact knowledge about his pupils.
- 6 If standardized test results are to be used in measuring the efficiency of instruction, the conditions of scientific experimentation must prevail with *contributing factors defined, measured, and controlled*. Failure to observe these conditions often results in teaching for test results alone, which not only invalidates any results which may be obtained, but also neglects some of the most desirable outcomes of good teaching which cannot be measured by tests. On the other hand, standardized test results cannot be ignored. They can be of great help to an administrator in judging a teacher's work, but they cannot be used as a substitute for classroom visiting, supervision, and critical subjective analysis.
- 7 No important decision regarding the placement of an individual pupil should be made on the basis of the result of one test of any kind. Educational achievement, mental age, I.Q., chronological age, health, teacher's judgment, physical development, social age, and emotional maturity are all factors to be considered in individual placement or any plan for grouping.

⁴ Arthur E. Traxler, *Planning and Administering a Testing Program*, *School Review* 48: 253-267, April 1940.

⁵ *The Use of Standardized Tests in Massachusetts*, *Test Service Bulletin* No. 38, published by World Book Company, 1938.

- 8 The content or items of a standardized test should never be used as material for class presentation and drill either before or after the administration of the test. To reproduce any part of the test, either on paper or the blackboard, is not only a violation of the publisher's copyright, but will invalidate that test for future use in the school. For this reason, all copies of standardized tests should be accounted for, and extra copies should not ordinarily be left in the hands of the classroom teacher.
- 9 The I Q or mental age obtained from one group test of intelligence is less reliable than an average of I Q 's or mental ages obtained from the results of two or more group tests of intelligence. An individual test of intelligence is more valid and reliable than group tests only when it is administered by a skillful and well trained psychometrist.
- 10 The use of standardized tests and a knowledge of the methods used in their construction should result in an improvement in teacher-made measures of achievement.

One must not assume that the testing program should be restricted to the use of standard tests. As has been explained in the three preceding chapters *informal or teacher-made tests will have a large place in any complete testing program*. Schools should have a carefully thought out general policy on such matters as the frequency of testing, the importance of final examinations, the factors to be considered in determining final marks, and, most important of all, the uses to be made of the results.

Regardless of its scope, the complete testing program at any particular time will ordinarily consist of the following eight steps, or stages, in chronological order:

- 1 Determining the purpose of the program
- 2 Selecting the appropriate test or tests
- 3 Administering the tests
- 4 Scoring the tests
- 5 Analyzing and interpreting the scores
- 6 Applying the results
- 7 Retesting to determine the success of the program
- 8 Making suitable records and reports

A. Determining the Purpose of the Program

It must be recognized at all times that tests are only tools, and that measurement is always a means to an end, never an end itself. In the final analysis, then, the value of any testing program depends upon the use made of results. Unless something is going to be done about it in the end, there is no point to beginning. Merely "giving tests" without rhyme, rule, or reason is money, time, and effort wasted. The author once heard an experienced educator say that he had wondered for years what many people did with standard tests after they had been "given." At last he found out. They filed them! The testing program should have a more serious purpose

than that. The first step, therefore, in planning a program is to determine its purpose. In so doing, three things should be kept in mind:

- 1 It should be co-operative
- 2 It should be practical
- 3 It should be definite

A co-operative program. As a rule, the program should not represent the judgment of any one person alone, but that of a group. It should be a truly co-operative enterprise. The teachers and administrative officers alike should be made to feel that it is "our" program, as, indeed, it should be. This is not likely to be the case, however, if the principal, superintendent, or research department determines the program and then "hands it down" to the classroom teachers. The entire staff should have a voice in determining the purpose of the program and in formulating the plans, and all should have the opportunity of participating in it in every way possible from beginning to end. If this is not done, the teachers are not likely to understand the program fully or to appreciate what it is attempting to do. Without the hearty co-operation of the entire staff, from the superintendent to the youngest teacher, the program is almost sure to fall short of its highest possibilities. It is suggested, therefore, that in the small school or school system the purpose of the program be decided upon after discussion in a general teachers' meeting or series of meetings in which everyone has a chance to participate. In the larger school systems it is better to entrust a committee representing all interested groups with the responsibility of planning the program. Even then it should be brought before the entire staff before final action is taken. It cannot be emphasized too strongly that the success of the program largely depends upon co-operative action. An important part of the program, therefore, is the educating of the staff so that they can participate intelligently in it. Boyer emphasizes the fact that the teacher's attitude is the most important factor to be considered in any plan for "what she thinks and what she does as a result of her thinking, determines the success or failure of the plan."⁶

A practical program. The general purpose of the testing program is to provide data which will help in the solution of some practical school problem. As a rule, this means that the problem whose solution is sought will have to do with administration, instruction, or research or with some combination of these three. Even when tests are used primarily for administrative purposes such as classification, they can also be used by the classroom teachers for diagnostic purposes. Unless the school has had considerable experience with testing it will be better not to undertake a program primarily for research, although under favorable conditions research

⁶ *Thirty-Fifth Yearbook of the National Society for the Study of Education, Part I*
page 213. Bloomington, Illinois: Public School Publishing Company, 1936.

is a legitimate interest both of classroom teachers and of administrators. Even when the program is undertaken for research purposes, it should ordinarily be one which bears directly upon some practical issue in the school, such as determining the relative efficiency of different teaching methods or of administrative organizations.

A definite program. It is not enough that the program be co-operative and practical. It must also be definite. The scope of the program may vary all the way from a single subject in one grade to a complete measurement of the entire school system. A common mistake of a staff inexperienced in the use of tests is to undertake too much. The danger then is that the program will drag along until everybody is more or less "fed-up" with it. Much of the value of the information sought from the tests will be lost unless the information is made available without delay. It is usually best, particularly with inexperienced teachers, to run the risk of undertaking too small a program rather than one too large.

Another mistake is in stating the purpose of the program in too general terms. "To improve instruction" is too vague and inclusive. "To motivate study" or "to diagnose weaknesses and provide a basis for remedial instruction" would be better. Best of all would be a still more definite formulation, such as "to motivate study in fifth-grade arithmetic" or "to make a diagnosis of characteristic weaknesses in first-year algebra and to formulate a program of remedial teaching to strengthen them." The purpose should state specifically both the *nature* and the *scope* of the program to be undertaken. Later chapters will discuss in some detail important administrative and instructional problems which tests may help to solve. In a long time program the purpose for each year will have a definite relationship to the whole. No matter how stated, however, there is really one fundamental purpose in all measurement—namely, the better understanding of the individual pupil. To accomplish this purpose the information must be as definite and as complete as possible.

B. Selecting the Appropriate Test or Tests

When the purpose of the testing program has been determined, and not until then, the selection of the test, or tests, is in order. In Chapter 4 attention was called to the fact that a test may be superior for one purpose and worthless for another. Great care must therefore be exercised in order to secure the tests most appropriate for the purpose. Three questions require consideration:

- 1 Who shall select the test or tests?
- 2 What type of tests shall be used?
- 3 What is the best procedure in making the selection?

Who shall select the tests? The best qualified person, or persons, available should make the selection. In larger school systems the director

of research is usually that person. But, even then, in the selection of achievement tests for specific subjects, the teachers of these subjects should be consulted, as their knowledge is essential in judging the curricular validity of the tests. In smaller schools the major responsibility is usually entrusted to the principal or superintendent.⁷ However, in the selection of achievement tests a committee of teachers will be helpful in judging the content of the tests. It is a sound principle in all evaluation that involves a subjective element to rely, whenever possible, upon the combined judgment of a group of competent persons rather than upon the judgment of any one individual.

What type of tests shall be used? Ordinarily an adequate testing program will involve the use of more than one type of test. It will be desirable, except in a few cases such as in the beginning of the kindergarten or first grade, to use both intelligence and achievement tests. If considerations of time and money make it advisable to limit the testing program to one standard test for determining the present status of the class or school the best choice will usually be a test battery.⁸

For a general survey of the intellectual status of the class or school a good group test of intelligence will suffice, although as a rule an average of two is better than one alone. In any measurement of intelligence involving group tests, especially if only one test is used, it is desirable to have retested with an individual intelligence test, such as the Revised Stanford Binet or the Wechsler Bellevue, the following pupils: those who test very low, say below an IQ of 80, those who test very high, say above an IQ of 130, or those whose scores are considerably out of line with the judgment of the teacher. The Revised Stanford Binet is particularly trustworthy at the low IQ levels. The distinctive advantage of the individual intelligence test is the opportunity afforded for the examiner to observe the behavior of the child under standardized conditions. As a diagnostic instrument such a test is likely to be much superior to the group test. Pupils who have language difficulty should be tested individually, perhaps with a performance test.

A reasonably complete testing program will require as a rule the use of intelligence tests along with achievement tests. Because of the relative constancy of the IQ it is unnecessary to administer intelligence tests each year. The mental level of most pupils can be predicted closely enough from intelligence tests periodically scheduled to permit ordinary comparisons with achievement. Page 225 outlines intelligence testing programs adapted to various types of school organization. At times aptitude tests in specific fields, rating scales, check lists, personal interviews, and the like will also

⁷ State wide surveys in Massachusetts and New Jersey indicate this clearly. See *Test Service Bulletins No. 38 and No. 42*, World Book Company.

⁸ Nearly every major test publisher has such a battery. The interested administrator or other person responsible for helping with the planning of testing programs will want to have test catalogues of the companies listed with asterisks in Appendix F, page 464.

TABLE

ADVANTAGES AND LIMITATIONS OF STANDARDIZED

CRITERION	STANDARDIZED	
	<i>Advantages</i>	<i>Limitations</i>
1 Validity		
a Curricular	Careful selection by competent persons after experimentation Fit typical situations	Inflexible Too general in scope to meet local requirements fully especially in unusual situations
b Statistical	With best tests high	Criteria often defective Size of coefficients largely dependent upon range of ability in group tested
2 Reliability	With best tests very high usually above 90 often above 95 Usually fully objective	No guarantee of validity Depends upon range of ability in group tested
3 Usability		
a Ease of Administration	Definite procedure time limits etc Economy of time	Manuals require study and are sometimes inadequate
b Ease of Scoring	Definite rules keys etc Largely routine	May take considerable time Monotonous
c Ease of Interpretation	Better tests have adequate norms Useful basis of comparison Equivalent forms	Norms often confused with standards Some norms defective Norms for various types of schools and levels of ability are often lacking
Summary Main Points Pro and Con	Convenience comparability objectivity Equivalent forms may be available	Inflexibility

be required The particular combination of measuring techniques required in any given situation will depend upon the specific purposes to be served As a rule, classroom teachers will find a larger place for nonstandardized teacher made tests in the solution of instructional problems than will school administrators in the solution of administrative problems The re-

AND NONSTANDARDIZED TESTS OF ACHIEVEMENT

NONSTANDARDIZED			
ESSAY		OBJECTIVE	
<i>Advantages</i>	<i>Limitations</i>	<i>Advantages</i>	<i>Limitations</i>
Useful for English-advanced classes afford language training May encourage sound study habit*	Limited sampling Bluffing is possible Mix language factor in all scores Usually not known	Extensive sampling of subject matter Flexible in use Discourage bluffing Easier to prevent and to detect cheating Compare* favorably with standard tests	Narrow sampling of functions tested Negative learning possible Piecemeal study encouraged Adequate criteria usually lacking
Inexperienced teachers may do better than with objective types	Average is low Subjective scoring	Sometimes approaches that of standard tests Objective scoring	No guarantee of validity
Easy to prepare Easy to give	Lack of uniformity Slow uncertain and subjective No norms Meaning doubtful	Directions rather uniform Economy of time Definite rules keys etc Largely routine Local norms can be derived	Difficult to prepare May take considerable time Monotonous No norms available at beginning
Useful for part of many tests and in a few special fields	Limited sampling Subjective scoring Time consuming	Extensive sampling Objective scoring Flexibility	Preparation requires skill and time

verse condition will tend to be true for standardized tests Table 27 is a sort of "balance sheet" which briefly summarizes some of the chief advantages and limitations of various types of achievement tests It is evident that there is a legitimate place for all kinds of tests, but no one test is equally good for all purposes

What is the best procedure? Regardless of the purpose of the testing program or who makes the selection of tests, it is important that a systematic, businesslike procedure be employed. Users of standard tests will find the information contained in *The Mental Measurements Yearbooks* of great value. The comprehensive character of the tests reviewed in this publica-

TABLE 28

CLASSIFICATION OF TESTS IN *The Fourth Mental Measurements Yearbook*
(1953)

TEST	TEST
ACHIEVEMENT BATTERIES.. 1	SAFETY EDUCATION..... 521
CHARACTER AND PERSON- ALITY	TESTING PROGRAMS..... 526
NONPROJECTIVE..... 27	READING..... 528
PROJECTIVE..... 102	MISCELLANEOUS..... 561
ENGLISH..... 148	ORAL..... 565
COMPOSITION..... 178	READINESS..... 566
LITERATURE..... 180	SPECIAL FIELDS..... 573
SPELLING..... 198	STUDY SKILLS..... 578
VOCABULARY..... 213	SCIENCE..... 589
FINE ARTS	BIOLOGY..... 596
ART..... 219	CHEMISTRY..... 607
MUSIC..... 225	GENERAL SCIENCE..... 623
FOREIGN LANGUAGES..... 232	GEOLOGY..... 630
ENGLISH..... 233	MISCELLANEOUS..... 631
FRENCH..... 236	PHYSICS..... 633
GERMAN..... 244	SENSORY-MOTOR..... 644
GREEK..... 248	HEARING..... 646
ITALIAN..... 249	MOTOR..... 649
LATIN..... 250	VISION..... 654
SPANISH..... 259	SOCIAL STUDIES..... 662
INTELLIGENCE	ECONOMICS..... 670
GROUP..... 267	GEOGRAPHY..... 674
INDIVIDUAL..... 334	HISTORY..... 679
MATHEMATICS..... 365	POLITICAL SCIENCE..... 698
ALGEBRA..... 380	SOCIOLOGY..... 708
ARITHMETIC..... 398	VOCATIONS..... 710
GEOMETRY..... 422	CLERICAL..... 719
TRIGONOMETRY..... 438	INTERESTS..... 736
MISCELLANEOUS	MANUAL DEXTERITY..... 749
AGRICULTURE..... 441	MECHANICAL ABILITY..... 756
BUSINESS EDUCATION..... 443	MISCELLANEOUS..... 777
COMPUTATIONAL AND SCORING DEVICES..... 464	SPECIFIC VOCATIONS..... 785
ETIQUETTE..... 471	ACCOUNTING..... 787
HANDWRITING..... 475	DENTISTRY..... 788
HEALTH..... 478	DRIVERS..... 789
HOME ECONOMICS..... 491	EDUCATION..... 792
INDUSTRIAL ARTS..... 503	ENGINEERING..... 808
PHILOSOPHY..... 505	LAW..... 814
PSYCHOLOGY..... 507	MACHINISTS..... 816
RECORD AND REPORT FORMS... 510	MEDICINE..... 817
RELIGIOUS EDUCATION..... 518	NURSING..... 818
	SALESMEN..... 824

tion is indicated by the "Classification of Tests" in *The Fourth Mental Measurements Yearbook*, which is shown in Table 28.⁹

As an illustration of the type of evaluations in this volume the following excerpts from comments on the Primary Mental Abilities tests are given.¹⁰

Anne Anastasi, Professor of Psychology at Fordham University, criticizes the PMA tests on the basis of the methods used to estimate their reliability

A special weakness of the entire PMA series is the treatment of test reliability. In tests such as these designed for intra individual comparisons and profile analysis the need for proper determination and reporting of reliability is particularly urgent. Yet in the various forms of the PMA tests reliability coefficients are either inadequately reported, incorrectly computed, or completely omitted. Odd-even and Kuder Richardson techniques have been repeatedly employed in finding the reliability of speeded tests [for which they are not suitable]. In several forms no recognition is given to this problem at all; spurious and meaningless reliabilities as high as .95 being reported without comment except to say that the reliabilities would probably be higher in more heterogeneous samples.

Ralph F. Berdie, Professor of Psychology and Director of the Student Counseling Bureau at the University of Minnesota, does not deliver a favorable final verdict.

In general one would expect these tests to be a great contribution to education and guidance. That they have not been may be due either to the test itself or to the inadequate follow up work that the authors or others have done. It may be that in attempting to produce a test that requires relatively little time or money the publishers have sacrificed those very things that made the tests potentially valuable. It is too bad that after such tests have been available for more than 14 years, one must still conclude that their principal uses are experimental.

John B. Carroll, Associate Professor of Education at Harvard University, is more complimentary.

The authors are undoubtedly on solid ground in their discussions of the Verbal Meaning and Reasoning factors. In all probability the statements which the authors make about the Number and the Space factors and their relevance to certain types of curricula and jobs will eventually be substantiated in validity studies; but this is only the reviewer's hunch.

Stuart A. Courtis, Professor Emeritus of Education at the University of Michigan, starts off flatteringly but ends by questioning the value of all tests.

No tests this reviewer has ever seen or used approach the PMA tests in the care and ingenuity evident in their construction. The authors have very wisely broken away from conventional memory question response type of items. In all tests the exercises involve mental functioning in action. In other words the tests and manuals might well serve as models for all publishers of tests to follow. This reviewer, however, rejects as inappropriate or totally false both the statistical meth-

⁹ Oscar K. Buros (Editor) *The Fourth Mental Measurements Yearbook* page vii Highland Park, New Jersey: The Gryphon Press, 1963.

¹⁰ *Ibid.* pages 698-710.

TABLE 29

OTIS SCALE FOR RATING STANDARD TESTS¹¹

Scale for Rating Tests	Names of Tests				
Manual (5)					
Validity (15)					
Reliability (10)					
Reputation (5)					
Ease of Administration (Total 15)					
(a) Preparation (4)					
(b) Time limits (4)					
(c) Explanation needed (3)					
(d) Alternative forms (4)					
Ease of Scoring (Total 15)					
(a) Objectivity (10)					
(b) Time required (3)					
(c) Simplicity (2)					
Ease of Interpretation (Total 15)					
(a) Norms (5)					
(b) Directions for interpreting (4)					
(c) Class record (1)					
(d) Application of results (5)					
Convenient Packages (5)					
Typography and Makeup (5)					
Test Service (10)					
Total (100)					

ods and the theories of primary mental abilities derived from their use. This reviewer predicts that the PMA tests, in spite of their structural and functional excellence, will not yield laws or educational principles any more than other tests have done.

¹¹ Published by World Book Company, Yonkers, New York.

P. E. Vernon, Professor of Educational Psychology in the Institute of Education of the University of London, seems favorably impressed. He points out that:

Thurstone has clearly retreated from his earlier opposition to "general" intelligence. He not only allows total scores to be calculated for each battery, but even provides for their conversion to IQ's.

Thus there are reviews of the PMA series by five different persons which fill ten large double-column pages and together cover almost all points of interest to nearly any prospective user of the tests. Few of the other tests get this much coverage, many are reviewed by only a single person.

The *Fourth Yearbook* also cites nine reviews of PMA tests in previous yearbooks.

In the choice of standard tests it is always wise to have available for careful examination both the test blanks and the test manuals of all tests being considered. Most county and city school systems will find it desirable to maintain for such purposes up-to-date sample ("specimen") sets of the more important tests published. To assist in making the necessary examinations and comparisons, the use of a rating scale will be found helpful. The first one published, prepared by Otis, is reproduced in Table 29. A more analytical scale for evaluating achievement tests is that of Cole and von Borghersrode, given in Table 30.

The use of these scales not only directs attention to significant points but also gives some idea of the relative weight of the various items. In the authors' opinion, Cole and von Borghersrode assign too much weight to reliability, and both scales assign too little weight to validity, the most important quality of any measuring instrument. Also, the relative weight assigned by both these scales to what may be termed *usability* seems somewhat heavy. The authors suggest a slight revision in weights and a regrouping of sections IV, V, VI, and VII under the heading *usability*. The major divisions and subdivisions, with revised weightings, would then be as follows:

<i>Division</i>	<i>Points</i>
I Preliminary Information	
II Validity	50
A Curricular	30
B Statistical	20
III Reliability	20
A More important points	15
B Less important points	5
IV Usability	30
A Ease of administration	10
B Ease of scoring	5
C Ease of interpretation	10
D Miscellaneous	5
Total	100

TABLE 30

COLE-VON BORGERSRODE SCALE FOR RATING STANDARDIZED TESTS¹²**I Preliminary Information**

- 1 Exact name of test
- 2 Name and position of author
- 3 Name of publisher and nearest address
- 4 Cost
- 5 Date of copyright
- 6 Purpose of test

II Validity (20)**A. Curricular (15)**

- 1 Exact field or range of educational functions which test measures?
- 2 Ages and grades for which intended?
- 3 Criteria with which material was correlated?
- 4 Do questions parallel good teaching procedures?
- 5 How wide is sampling of important topics?
- 6 What is the social utility of questions?
- 7 Is test claimed to be diagnostic? (If so, see VI, 5, c, below)

B Statistical (10)

- 1 Correlated against what outside criteria?
- 2 Size of coefficient of correlation?
- 3 Size and representativeness of sampling?
- 4 Proof of adequacy of items (such as statements as to experimental try out of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels)

III Reliability (25)**A. Most important points**

- 1 Correlated with what?
- 2 Size and representativeness of sampling?
- 3 Reliability coefficients
- 4 The means of the distributions
- 5 The standard deviations of the distributions
- 6 Other similar statistics
- 7 Intercorrelations

B Less important but desirable

- 1 Order of giving various forms of test
- 2 Is test reliable enough statistically for individual measurement, or should it be used only for groups?
- 3 Evenness of scaling (see II, B, 4)
- 4 Are pupils accustomed to this type of test?

IV Ease of Administration (15)**1 Manual of Directions (3)**

- a How complete and simple is the manual?
- b Does manual control test conditions well?
- c Typographic makeup

¹² Robert D. Cole and Fred von Borgersrode, 'A Scale for Rating Standardized Tests' *School of Education Record of the University of North Dakota*, 14: 11-15, October 1928.

TABLE 30 (Continued)

COLEMAN BORGENSRODE SCALE FOR RATING STANDARDIZED TESTS¹²

IV Ease of Administration (15) (Cont)

- 2 Simplicity of administration (9)
 - a Amount of explanation needed for pupils by examiner?
 - b Are directions to pupils clear detailed comprehensive?
 - c Is arrangement of test convenient for pupils?
 - d Are samples and fore-exercises given when needed?
 - e Time needed for giving?
- 3 Alternate forms (3)
 - a Number
 - b Evidence of reliability
 - c Evidence of equivalency

V Ease of Scoring (10)

- 1 Degree of objectivity—purely objective or some judgment on part of examiner?
- 2 Are adequate directions given—clear equal to all emergencies?
- 3 Is scoring key adjusted to size of test?
- 4 Time needed to score one test
- 5 Simplicity of procedure
 - a Number of processes needed to get final score?

VI Ease of Interpretation (20)

- 1 Norms (6)
 - a Kind—age grade percentile standard score etc
 - b Derivation—size and representativeness of sampling
 - c Tentative arbitrary or experimental?
 - d For separate parts?
 - e How expressed?
- 2 Is class record provided? (1)
- 3 Are there provisions for graphing results? (1)
- 4 Is interpretation of scores easy or hard? (2)
- 5 Application of results (10)
 - a. Are directions or suggestions given for application of results to benefit teaching or administration?
 - b Are tests survey or diagnostic?
 - c If diagnostic—
 - (1) Proof of diagnostic value?
 - (2) What principle or principles underlie construction?
 - (3) How many different skills abilities or aspects of the subject are analyzed or measured?
 - (4) Does the analysis of total subjects into unit abilities follow teaching practices?
 - (5) Is the diagnosis individual or class—proof?
 - (6) Does the test demand tabulations of individual pupils errors to secure a diagnosis?
 - (7) Is a remedial program provided or suggested?

VII Miscellaneous (5)

- 1 Typography and makeup
 - a Arrangement of printed matter
 - b Legibility of type
 - c Quality of paper
 - d Are test blanks free from distractions norms directions to examiner etc?

TABLE 30 (Continued)

COLE-VON BOGERSPOFF SCALE FOR RATING STANDARDIZED TESTS

VII Miscellaneous (5) (Cont.)

- 2 Is the time required for giving as small as is consistent with reliable measurement?
- 3 Is the cost in keeping with the amount, scope, and reliability of the results yielded?
- 4 Is good test service provided by the publisher?
- 5 Kind of objective questions used?

A desirable procedure is to have a group of at least three competent people, each independent of the others, look over all the tests being considered, the manuals accompanying them, and any evaluations available. Each judge first compares the tests with respect to validity, and records the judgment in points before considering anything else. Then he goes on to reliability and makes a similar judgment on each test. Finally, he does the same for usability. This method will tend to produce greater agreement among the judges regarding the *relative ranks* of the tests on the criteria individually. After all, the total point score allowed a test is less important than the rating on the divisions separately.

Emphasizing the close relationship between teaching and testing, Brownell suggests the following criteria¹² for evaluating tests:

- 1 Does the test elicit from the pupils the desired types of mental processes?
- 2 Does the test enable the teacher to observe and analyze the thought processes which lie back of the pupils' answers?
- 3 Does the test encourage the development of desirable study habits?
- 4 Does the test lead to improved instructional practice?
- 5 Does the test foster wholesome relationships between teacher and pupils?

In selecting a test for a given purpose, the grade level on which it is to be used must be given consideration. Test publishers often suggest a considerable grade range in which the test may be used. But both test authors and publishers tend to be too optimistic concerning the range of usefulness of their tests. For example, an intelligence test that is supposed to be suitable for grades three to eight may be found to be too difficult for the third grade and too easy for the eighth. The reader will doubtless recall from a discussion in Chapter 4 that it has usually been found that a test has optimum discrimination for a group whose average corrected-for-chance score is approximately 50 per cent of the maximum score possible on the test. It must be remembered, however, that the discriminating function of diagnostic and certain other specific tests is usually relatively unimportant.

¹² William A. Brownell, "Some Neglected Criteria for Evaluating Classroom Tests," *National Elementary Principal* 16: 485-492, July, 1937.

C. Administering the Tests

The next step in the testing program is the administering of the tests. In order to insure that this is properly done, three questions must be answered

- 1 When should the tests be administered?
- 2 Who should administer the tests?
- 3 What is the correct procedure to follow?

Each of these questions deserves careful consideration

When should the tests be administered? As problems concerning the use of intelligence tests differ somewhat from those concerning the use of achievement tests alone, it is better to consider the two separately. When should intelligence tests be administered? There is general agreement that it is not necessary to give the same pupils intelligence tests every year, but there is also agreement that possible fluctuations on group tests are great enough to warrant giving such tests more than once. The fluctuations are likely to be most serious in the primary grades.¹⁴ A reasonable plan employed by many school systems is to give intelligence tests at transitional points in the pupil's school history. As Stoddard suggests, "Intelligence is analogous to health, any estimate of it should be rechecked close to the making of an important decision."¹⁵ Procedure would therefore vary according to the school organization. A suggested minimum program is as follows:

Type of Organization	Grades to Give Intelligence Tests
Six six plan	First and sixth or seventh
Seven five plan	First and seventh or eighth
Eight four plan	First and eighth or ninth
Six three-three plan	First, sixth or seventh, and ninth or tenth

If possible, it would be well to add to this minimum program a test at about the fourth grade and one at the end of the high school course.

There is some disagreement regarding the best time of year in which to give the intelligence tests. Of course, if the tests are to have maximum value, their results must be made available at the very beginning of these transitional periods. This means they should be given early in the first grade if the pupils have had no previous kindergarten experience. Since Updegraff¹⁶ found that for preschool children the reliability of the test is

¹⁴ Cf. Mildred M. Allen, "Relationship between the Indices of Intelligence Derived from Kuhlmann Anderson Intelligence Tests for Grade I and the Same Tests for Grade IV," *Journal of Educational Psychology* 36: 252-256, April 1945.

¹⁵ George D. Stoddard, *The Meaning of Intelligence*, page 94. New York: The Macmillan Company, 1943.

¹⁶ Ruth Updegraff, "The Determination of a Reliable Intelligence Quotient for the Young Child," *Pedagogical Seminary and Journal of Genetic Psychology* 41: 152-166, September 1932.

increased by postponing testing until two weeks after entrance to school, it may be well to avoid giving the test till the second or third week of school in the lower grades. The later tests can be given either at the beginning of the transitional year or at the close of the year preceding. There is a tendency to have tests for college entrance administered in the high schools near the close of the senior year. This is obviously necessary if such tests are to be used in counseling these seniors regarding the feasibility of continuing their education. There will usually be a few pupils who will transfer into the system and who have not had intelligence tests, and others in the system about whom teachers may feel serious doubt regarding the validity of the existing record.

The frequency with which achievement tests should be used will depend primarily upon the purpose they are to serve. Most purposes, however, will require at least two series of tests administered at intervals of a semester or a year. Most achievement tests have norms for the middle and the end of the year, but often for no other time. When tests are given at these periods, comparisons with norms are easiest. There is also the fact that many studies have shown a considerable decline in knowledge at the end of the summer vacation. This would seem to favor giving the tests at the end of the school year, when the pupils' status is more normal. A comparison between the records made by pupils at the end of each of two successive years is usually more trustworthy than that between the beginning and end of one year.

There are some advantages in having the tests administered in the fall. Almost always some pupils will enter the school for the first time and their status can best be determined by administering tests to all the pupils. The teachers will then have the entire school year in which to remedy any deficiencies revealed. Fall testing also avoids the undesirable practice of cramming. If too much emphasis is placed on "improvement" shown during the year, however, pupils may be tempted not to do their best on the first series of tests. This would not be the case if progress is measured between two series of tests administered at the end of the preceding year and at the end of the current school year.

This practice will also make it possible to have the information serve several purposes. It can be used partially as a basis for determining promotion from the grade, for educational guidance and possibly for sectioning the next grade. There seems also no good reason why an analysis of the errors revealed cannot serve equally well as a basis for remedial teaching in the succeeding grade as if the new teacher had given the test at the beginning of the year. Of course, in some instances there might be considerable value in repeating the test at the beginning of the year in order to determine the effects of the summer vacation, apart from the better established weaknesses which were present when the vacation started.

Moreover, the analysis of errors is more trustworthy when based upon two samplings of performance than upon one

Who should administer the test? Obviously, only competent persons should administer standardized tests. It is not always an easy matter to tell who is really competent, however. In the case of individual tests of the Stanford-Binet type, this requirement means that only persons who have had specific instruction in college classes should attempt to administer them. There should be at least one person in every school who is qualified to give such tests. When tests are used for purposes of research, or when they are used to compare one grade, class, or school with others, they should usually be given by one person, or a small group of specially trained examiners. But in the ordinary testing program, employing group intelligence tests and achievement tests, the regular classroom teachers should usually administer the tests. Most of them will welcome an opportunity to do so. At the present time there seems no good reason for selecting a test whose administration is so difficult as to be beyond the mastery of average teachers in the public schools. The point of view of McCall seems eminently sound:¹⁷

Many years ago certain specialists sought to secure a monopoly of the privilege of using standard tests by trying to persuade educators to regard the tests as possessing certain mystic properties. A few of us with Promethean tendencies set about taking these sacred cows away from the gods and giving them to mortals. Can teachers be entrusted with tests? If not, then teachers ought not to be trusted with 90 per cent of their present functions. We now entrust them with the far more difficult task of teaching reading, creating concepts and building ideals. Let us not strain at a gnat when we have swallowed fifty elephants.

But it is well not to take the competency of the examiners for granted. One of the best plans is to get the group of examiners together and demonstrate the administration of the tests to be used. One way to do this is to give a demonstration with a regular class and to follow this by a discussion with the examiners of the procedure they have seen. Another way is to administer the test to the examiners themselves. This should be followed by a full discussion of the procedure involved. It is usually well to suggest that after each examiner has studied the manual he try the procedure on some other person, such as a member of the family, or two teachers may try it out on each other. If questions then arise they can be settled by a conference with the person in general charge of the program before the examiner goes before his group actually to administer the test. It has been found that, if such measures are taken, the regular classroom teachers can obtain practically the same results with group tests as can be obtained by special examiners.

¹⁷ W. A. McCall in *The Test Newsletter*, published by Bureau of Publications, Teachers College, Columbia University, December 1938.

What procedure should be followed? Although the procedure of administering group intelligence tests and achievement tests is not beyond the mastery of classroom teachers and school administrators, some difficulties may arise. In fact Ligon¹⁸ argues that good group testing is more difficult than individual testing. In the first place, the conditions for the test must be favorable. It is usually best to have the tests given in the familiar environment of the pupils' own classrooms. Especially is this true of younger children. It is well always to have the tests given at regular class time without permitting them to run over into lunch hour or play time. For the same reason it is desirable not to have tests just before or just after an important event, such as a holiday, a school party, or an athletic contest. Precautions should be taken to avoid all unnecessary distractions and interruptions during the progress of the test. It is a good plan to hang on the outside of the classroom door a card which reads *Tests Going On Please Do Not Disturb*. Pupils should be instructed to remove everything from the tops of their desks except two well-sharpened pencils and an eraser. The examiner should also have ready a few extra pencils in case of an emergency. All these things must be looked after in order to insure favorable working conditions for the test.

As a rule, anyone can administer a group test successfully who meets three requirements. The first of these is the ability to read well. Good silent reading is required for the mastery of the directions printed in the manual which accompanies the test. Good oral reading ability is required, for the directions to the pupils should be *read*, not recited from memory. To undertake to give the test from memory is to run a serious risk of leaving out some important word or phrase or of paraphrasing the directions in such a way as to change their meaning. But the examiner should be so familiar with the manual that he can read the directions with his eyes off the page a good part of the time. The directions should be read with proper emphasis in a clear voice just loud enough to be heard throughout the room. The aim should be to make the meaning understood without arousing anxiety or excitement.

The second requirement for administering a test is the ability to keep time accurately. If the test has a single time limit of, say, twenty minutes or more, it is probably preferable to time it with an ordinary pocket watch rather than a stop watch, since the latter may, upon occasion, be quite erratic. When a pocket watch is used, set its hands to some convenient time such as the beginning of an hour and give the starting signal just as the second hand reaches 60 (which is also 0). It will usually help students and examiner alike to have a clock in the room which shows everyone the correct time. Some testers use a special device known as an interval timer.

The aim should be to keep the time to a second. On most tests the signal

¹⁸ Ernest M. Ligon. *The Administration of Group Tests*. *Educational and Psychological Measurement* 2: 387-399. October, 1942.

to start is, "Ready, go" or "Ready, begin!" When this signal is given the examiner should note the *exact* time—hour, minute, and second. This should be recorded *immediately* preferably on a small card or specially prepared blank. The record for Test 1 would look like this

Test 1			
Time test began	Hr	Min	Sec
Time allowed	9	0	0
		5	
Time to stop	9	5	0

Experienced examiners know that it is never safe to trust one's memory to keep the time. A written record must be made.

The third requirement for administering a test is the ability to follow directions accurately. The manual should be followed verbatim. No deviation whatsoever is permissible. To add anything to or to modify the directions in any way means that it is no longer a standardized test. Boynton¹⁹ gives some interesting illustrations of unconscious clues given by inexperienced examiners using the Stanford Binet. One examiner, for example, when asking the meaning of the word "tap" in the vocabulary test began to tap on the table and when he came to the word "eyelash" he looked the child straight in the eye and batted his eyes rapidly. The norms are made on the assumption that a prescribed formula is to be used. As a part of the preliminary instructions pupils are almost always told not to ask any questions after the test starts. Occasionally a pupil forgets this instruction and holds up his hand for a question. The examiner should walk over to him and, if it is a reading test or an intelligence test whose purpose is following directions, should say in a quiet voice, "Read it carefully and do just what it says." If it is an ordinary achievement test and the pupil is concerned about where to put his answer or some other point of mechanics that does not involve the answer to a question in the test or modify the directions already given it is permissible to set the pupil at ease without causing disturbance. Kelley suggests this principle in handling the child who is in trouble. "The examiner should be free to say or do anything that does not disturb or delay pupils at work, that does not help the individual child in the thing in which he is being tested, and that does set him to work again after some foolish or trivial issue has troubled him."²⁰ Examples of permissible statements are "Yes, you may change your response if you decide it is wrong," "Just work on the side of the sheet, you do not need scratch paper," "When you have finished the first column go right on to the next one," "No, you must not go back to a test you have passed" and the like. But if the pupil asks the meaning or spelling of a word, or how to answer

¹⁹ Paul L. Boynton *Intelligence Its Manifestation and Measurement* pages 276-277 New York: D. Appleton Century Company, 1933.

²⁰ Truman Lee Kelley *Interpretation of Educational Measurements* page 46 Yonkers World Book Company, 1927.

a test item, the examiner should say quietly "I cannot tell you Go on to the next one" *In case of doubt, the examiner should err on the side of saying nothing* While the test is in progress the examiner must be alert constantly to see that the pupils neither help nor hinder each other nor are distracted by external factors Ligon²¹ indicates the following requirements of good group testing "That all the subjects understand the instructions, that they all work throughout the assigned time at their optimum level of achievement, that they are in no way helped, hindered, or distracted by one another, that they do not quit trying or omit any section of the test, that examiners give instructions adequately and in a stimulating, effective tone of voice—not a dull bored monotone—and that proctors are observing every movement of the group, stimulating lagging souls, inhibiting wandering eyes, and detecting failure to follow instructions" A test is more than a measuring device, it presents a standardized situation in which to observe pupil behavior Any occurrence observed during the progress of the test that may throw light upon the interpretation of the results should be carefully recorded

D. Scoring the Tests

It is desirable to have the tests scored as quickly as possible and with the highest possible degree of accuracy As a rule, then, that system is best which accomplishes these objectives with the minimum expenditure of money, time, and energy There are two questions involved

- 1 Who should score the tests?
- 2 What technique should be used?

Who should score the tests? In actual practice, standard tests are scored by a variety of persons Sometimes, especially in larger systems, the work is done by a clerical staff at a central bureau, or by the use of scoring machines, the scoring may be contracted for with some outside agency, such as the test publisher, sometimes it is done by advanced students under supervision, at other times the scoring is done by administrative officials, but the most common method seems to be to have the work done by the regular teachers Except in the larger systems where there is a bureau of research equipped with special facilities, the scoring is probably best done by the classroom teachers In that way not only can the work be done promptly, but the teachers can probably learn something of value about the types of errors made on the achievement tests But it is important to get the scoring done without producing an unfavorable attitude toward it on the part of the teachers Some schools have found it very satisfactory to dismiss classes at noon when the testing is in progress, so that the teachers can devote the afternoon to the work of scoring This would seem

²¹ Ernest M Ligon *op cit* page 387

an effective way of emphasizing the important fact that teaching and testing are processes that are intimately related.

What techniques should be used? Every reasonable precaution should be taken to assure a high degree of accuracy in scoring. It must not be assumed that merely because the directions are clear, the key complete, the separate answer sheets well designed and the process entirely objective, perfect protection against errors is thereby afforded. Numerous studies give abundant evidence to contradict this assumption. They reveal two distinct types of errors in scoring: *constant errors* and *variable errors*. A common example of the former type is misunderstanding the scoring directions, for instance, by counting omissions the same as errors, when using the scoring or correction formula. Such errors are especially serious, because there is no possibility of their offsetting each other according to any so-called "law of averages." Variable errors, on the other hand, sometimes tend to make the score too high and at other times too low. While such errors may do serious harm to individual pupils, they tend to cancel each other in group measures such as averages. Examples of variable errors are errors resulting from carelessness, errors in counting the scores, errors in entering the scores on the front of the test booklet or on the record sheet, and errors in adding up the total score. Some of the most serious errors found are not in marking the paper at all but in counting and in addition.

Clearly, then, accuracy in scoring cannot be taken for granted. What is to be done about it? The first thing is to prevent the occurrence of errors whenever possible. The scorers must be taught how to score the papers and not merely told how to do it. They should be given an opportunity to study the manual and the scoring keys. Whenever possible, an actual demonstration of scoring should follow. It is a good idea, also, to check carefully the first few papers marked by beginners to detect errors at the outset. This procedure should reveal any constant errors and the principal types of variable errors. It is always desirable to have each page or part of the test scored through all the papers in a set before going on to the second page or part of the test. If the scorers work in groups, as is usually desirable, each one can specialize in marking one part of the test, and pass the test when scored to the next scorer, who is specializing in marking the next part of the test. This procedure will reduce the risk of error and at the same time will increase the speed of scoring. It is usually an especially poor technique to have one person read the answers while the scorers mark the papers. This is slow, because the slowest scorer sets the pace. It also increases the risk of error, owing to the possibilities of losing the place or of failure to hear correctly. Colored pencils are desirable. Inexperienced scorers should mark each item in the test being scored in some uniform manner, such as + for correct, - for incorrect, and 0 for omitted items. Experienced scorers will save time by marking only the incorrect and omitted items. It is of

TEST 3. WORD MEANING

When two words mean the SAME, draw a line under "SAME."
When they mean the OPPOSITE, draw a line under "OPPOSITE."

SAMPLES	fall — drop	same — opposite	
	north — south	same — <u>opposite</u>	
1	expel — retain	same — <u>opposite</u>	1 +
2	comfort — console	<u>same</u> — opposite	2 +
3	waste — conserve	same — <u>opposite</u>	3 +
4	monotony — variety	same — <u>opposite</u>	4 +
5	quell — subdue	<u>same</u> — opposite	5 +
6	major — minor	same — opposite	6 —
7	boldness — audacity	same — opposite	7 +
8	exult — rejoice	same — <u>opposite</u>	8 —
9	prohibit — allow	same — opposite	9 0
10	debase — degrade	<u>same</u> — opposite	10 +
11	recline — stand	same — <u>opposite</u>	11 +
12	approve — veto	same — <u>opposite</u>	12 +
13	amateur — expert	<u>same</u> — opposite	13 —
14	evade — shun	<u>same</u> — opposite	14 +
15	tart — acid	same — <u>opposite</u>	15 —
16	concede — deny	same — opposite	16 0
17	tonic — stimulant	<u>same</u> — opposite	17 +
18	incite — quell	<u>same</u> — opposite	18 —
19	economy — frugality	<u>same</u> — opposite	19 +
20	rash — prudent	same — <u>opposite</u>	20 +
21	obtuse — acute	same — opposite	21 0
22	transient — permanent	same — opposite	22 0
23	expel — eject	<u>same</u> — opposite	23 +
24	hoax — deception	<u>same</u> — opposite	24 0
25	docile — submissive	<u>same</u> — opposite	25 +
26	wax — wane	same — opposite	26
27	incite — instigate	same — opposite	27
28	reverence — veneration	same — opposite	28
29	asset — liability	same — opposite	29
30	appease — placate	same — opposite	30

Right...15...Wrong...5...Score...10...

Figure 6. An Illustration of the Procedure Followed in Scoring Test 3 of the Terman Group Test of Mental Ability, Form A. (Copyright by World Book Company.)

STANDARD TEST SCORING RECORD

Name of Test Madison School Achievement Test Form 1
 School Williamsdale Grade 4

No	Scored by	Errors	Checked by	Comment
1	<u>Tracy Anderson</u>	<u>5</u>	<u>John Long</u>	<u>In Adding</u>
2	<u>Julia Jones</u>	<u>0</u>	<u>Mabel Adams</u>	<u>Not reduced to</u>
3	<u>James Johnson</u>	<u>3</u>	<u>Lee Cross</u>	<u>lowest terms</u>
4	<u>James Long</u>	<u>4</u>	<u>Ed Howard</u>	<u>Decimals</u>
5	<u>James Lee</u>	<u>20</u>	<u>William Jamison</u>	<u>Misunderstood</u>
6	<u>Elma Wright</u>	<u>0</u>	<u>Jed Smith</u>	<u>Directions</u>
7	<u>Walter Justice</u>	<u>2</u>	<u>Yett Altus</u>	<u>Carelessness</u>
8	<u>Oscar Wilson</u>	<u>2</u>	<u>Ray Kelley</u>	<u>"</u>
9	<u>Hulda White</u>	<u>1</u>	<u>Jane Hunt</u>	<u>"</u>
10	<u>Anna Cargy</u>	<u>3</u>	<u>Henry Lister</u>	<u>"</u>

Transcribed by	Errors	Checked by	Comment
<u>Betty Brown</u>	<u>0</u>	<u>Walter Coleman</u>	
<u>Edith Raymond</u>	<u>1</u>	<u>Elizabeth Kent</u>	

Scores added by	Errors	Checked by	Comment
<u>Judson Allen</u>	<u>1</u>	<u>Alice James</u>	<u>Too low by 100</u>

Norms etc. by	Errors	Checked by	Comments
<u>Edna Harvey</u>	<u>0</u>	<u>Luan Clay</u>	

Class record by	Table made by	Median by	Graph by
<u>Martha Rule</u>	<u>Bert Tate</u>	<u>Norma Galt</u>	<u>Sarah Barr</u>

Figure 7 A Sample Standard Test Scoring Record

course, unnecessary to mark the items below the last one the pupil attempts. But it is well to draw a horizontal line across the test under the last item attempted. Figure 6 illustrates the scoring of an alternative-response test of word meaning, using the formula $\text{Score} = R - W$.

The writers have found that the simple device of keeping a written record of who marks, checks, transcribes, or totals each part of the test reduces the likelihood of error. If the scoring is organized systematically, it is a

simple matter to keep such a record on a mimeographed sheet attached to each package of tests when scored, as shown in Figure 7

But in spite of these preventive measures, certain errors are likely to occur. The safest plan, therefore, is to have each set of papers marked a second time by different scorers, using pencils of a different color. Dunlap²² found that items most subject to errors in scoring are of the two-response type requiring a scoring formula and items requiring the underlining of more than one word. If a complete rescoring does not seem practical, a sampling method may be followed. Each fifth or tenth paper, for example, may be selected and carefully rescored, and if only an occasional minor error is found, the whole set may be safely accepted. On the other hand, if frequent or serious errors are found in these sample papers, the entire set should be rescored. In any event it is important to have some person other than the original scorer check the totals for each part of the test and for the whole test, all substitutions in the scoring formulas, all transcribing of scores, and all transmuting of point scores into derived scores.²³ It is possible to locate many serious errors by examining closely the profile of each individual pupil on all tests with this form of record. Any score much higher or much lower than the general level is suspicious. Also, when two or more tests are used which purport to measure the same function, any serious discrepancies should be scrutinized, on the supposition that a high positive correlation is to be expected. The standard of absolute accuracy should be accepted by all scorers. *The possibilities of serious injustice to individual pupils by errors in scoring should be fully recognized.*

E Analyzing and Interpreting the Scores

After the tests have been scored and checked, the next step is the analysis and interpretation of the results. Both processes go on together, for analysis is worthless without interpretation and interpretation is impossible without analysis. Analysis is of two main types: statistical and graphical. Before either can be undertaken, however, there is the important preliminary step of classification and tabulation. An analysis of errors appearing in the test papers is usually of major importance to the classroom teacher. Chapters 3, 9, and 10 are concerned with a discussion of the whole problem of analysis and interpretation; only an outline will be given here to indicate the steps involved.

1. Classification and tabulation of scores
2. Statistical analysis of scores

²² Jack W. Dunlap, "The Relationship Between the Type of Question and Scoring Errors," *Journal of Experimental Education*, 6, 76-77, March, 1938.

²³ Derived scores are obtained from tables of norms. Each point score is expressed in some equivalent unit, such as an age or percentile score. The interpretation of these units is considered in Chapter 10.

- 3 Graphical analysis and representation
- 4 Use of norms and standards
- 5 Analysis of errors

In a complete testing program all five of these steps will receive attention, although not always to the same degree. If the primary purpose of the testing program is diagnosis, for example, the fourth step would be relatively unimportant and the fifth step relatively important. The reverse would be true of a program whose main objective is a study of the comparative efficiency of various grades, classes, and schools.

F. Applying the Results

The application of the results is the crux of the whole testing program. Everything that has gone before is really preliminary. Whatever value the tests are to have depends in the last analysis upon the use made of the results.

Just what is to be done, of course, depends upon the purpose of the program. Later chapters will consider in some detail the procedure to be followed for several administrative and instructional problems. It will be sufficient at this point to give some idea of how the procedure will vary with the purpose.

Suppose, for example, that the purpose of the tests is to determine the present status of a particular school with the idea of its improvement, and that the test data are before the principal. The question now is, what is to be done? Upon the basis of the test scores and other pertinent data such as the teachers' estimates, health reports, age-grade status and the like, several pupils are given trial promotions to the next higher grades. A small group of pupils, whose achievement and intelligence scores are well below the central tendency of their respective grades are organized into an ungraded class and put in charge of a teacher whose outstanding virtues are sympathy, patience, and common sense. Ability groups are also organized in a few grades and classes, with appropriate differentiation in curricula and methods.

Likewise, suppose the primary purpose of the testing program is to determine whether or not the teaching emphasis is correct in the various subjects in the grades and, when the test results are in, it is apparent that most of the grades are strong in arithmetic and spelling, about normal in reading, and weak in language and the social studies. Now what is to be done here? The principal calls the teachers together and presents the situation in tables and graphs, with suitable comments by way of interpretation. Then follows a regular "council of war." One or more committees are appointed to make a special study of the situation and to make recommendations at a meeting to be held a little later. Eventually, after discussion and deliberation, a course of action is decided upon, looking to the improvement of the situation in the weaker subjects.

The procedure will again be somewhat different in essential respects if the primary purpose is diagnosis and remedial work in reading. Here the test results should be analyzed in some detail in each grade. An analysis of the test papers, item by item, is often very revealing. Special effort should be made to locate the specific nature of the reading difficulties. There may be found some general weaknesses, such as the inability to use the index and table of contents in a book, or possibly to locate the central idea in a paragraph. There are usually, in addition, other weaknesses, which appear in certain pupils and not in others. Some of these will not be revealed at all by the usual paper and pencil reading tests, but will require special tools and techniques. After considering these facts, the staff will try to plan a remedial program to be followed during the year.

The essential point in all these cases is that *something is done about the situation revealed by the test scores*. To fail to apply the results in some practical way is to fail in the testing program.

G Retesting to Determine the Success of the Program

Most testing programs stop with applying the results, if, indeed, they go that far. But an essential step yet remains. After a reasonable time has been allowed for a trial of the remedial measures which were agreed upon in the light of the test data, a checkup should be made to determine the success of this program. Most tests are not sufficiently accurate to reveal progress over a shorter period than one half year. As a rule, a second form of the test or tests used in the beginning should be employed in retesting. If this is not done, it will usually be very difficult to express the results in terms sufficiently comparable to make an accurate measure of progress possible. Of course not all the gain found can be correctly attributed solely to the remedial program. Some of it is doubtless due to the practice effect or to familiarity with the test itself, part of it to teaching received outside the school, and part of it to natural growth. Often, however, the improvement will be so marked as to indicate beyond a reasonable doubt the effectiveness of the program attempted. At other times the improvement will be disappointingly small. It is then usually wise to modify the remedial program in the light of the results obtained.

The essential point is that the success of the remedial program must not be taken for granted. On the contrary, a definite effort must be made to check upon its effectiveness. To fail to do this is to leave the testing program incomplete. There is no better reason for taking the efficiency of the remedial program on faith than there was for taking the earlier results of teaching on faith.

H Making Suitable Records and Reports

Certain records and reports are essential to the success of the testing program. But by no means do all these records and reports come chrono-

logically at the end of the program. As a matter of fact, some of these are essential to the last three stages already discussed.

In general, it may be said that four groups have an interest in knowing what the tests show: the pupils, the teachers, the administrative officers, and the parents or public. The nature of the report will naturally vary somewhat with the group to whom it is made, and the nature of the record with the specific function it is to serve. However, regardless of the type of record or its specific function in any particular situation, its general function is always, as has been well stated by Stenquist, "to present test results and related information in such a meaningful way as to *arouse interest and action*, on the part of teachers, principals, supervisors, directors of special divisions, and superintendents."²¹

Report to pupils. The pupils have a right to know their performance on all achievement tests whether standardized or nonstandardized. In many cases it is well to go over the papers with the pupils in order to point out the nature of the errors made. The success of any remedial program will depend upon the pupils' cooperation. Long ago Thorndike stated the matter succinctly in these words:²²

The final justification for every testing regime rests in Mary Jones and John Smith, and it therefore behoves all persons who are making and giving tests to take them into partnership as soon and as completely as is feasible.

It is usually considered dangerous to present the results of intelligence tests to pupils. And there doubtless is more possibility of harm than of good in making known the mental ages and intelligence quotients of individual pupils. Difficulty is most likely to result from scores at the extremes of the distribution. Both pupils and parents can reconcile themselves to low scores on achievement tests, for that can be explained to their satisfaction on the ground that it is the school's fault. But low intelligence test scores seem to reflect directly upon the good name of the family, and this is resented. Only the exceptional pupil or parent has a fine enough philosophy of life to reconcile himself to the realities implied by a low score, and to resolve to make the most of it. There is also danger that the pupils with high test scores will be so inordinately puffed up as to endanger both their social standing with their fellows and their academic standing with their teachers. There are, however, special cases in which information regarding intelligence scores may properly be given. Some examples of these cases will be discussed in later chapters.

²¹ John L. Stenquist, "The Administration of a Program of Diagnosis and Remedial Instruction," *Thirty-fourth Yearbook of the National Society for the Study of Education*, page 518. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1935.

²² Edward L. Thorndike, "Tests and Their Uses," *Teachers College Record* 26: 93-94, October, 1924.

INFORMATION CONCERNING HOME										Date	Home Address	Lives With	Phone
Name		Nat'l	Sex	Dead	Religion	Language Spoken	Education						
Father													
Mother													
Step-Father													
Step-Mother													
Siblings		Older	Younger	Date	Father's Occupation	Mother's Occupation							
Brothers													
Sisters													
Others in Home													

[illegible][illegible]

Figure 9. Test Data Summary from the Cumulative Guidance Record of the Department of Supervision and Curriculum Development of the National Education Association

[illegible]

Figure 10. A Cumulative Record in Graphical Form.

consistency of progress made. A distinguishing feature of this record is the use of percentile ranks so spaced that equal distances vertically represent more nearly equal changes in achievement than do the percentile ranks themselves.

Such records, although easy to interpret, are somewhat laborious to prepare. Hagen²² found that the graphical record required twice as much time to prepare as the numerical record and was somewhat more subject to error. This is perhaps largely responsible for the discovery quite a few years ago that only about one third of the schools which held membership in the Educational Records Bureau made a graph of all test results.²³ Care must be exercised that such records do not become ends in themselves and that not so much time is devoted to their preparation that none remains for their practical use. Schools with limited resources and little clerical assistance should be content with less elaborate record systems than those which may be feasible for larger and wealthier schools.

Warnings concerning profiles. Because of their deceptive simplicity, profiles invite improper interpretations. Three precautions *must* be kept in mind.

1 Every point on an individual's profile should be based upon the same norm group as every other point, or at least upon highly similar groups. It is dangerous and misleading for example to have the percentile point representing the pupil's score on a clerical aptitude test determined by comparison with the scores of "employed males" while the percentile point for his mechanical aptitude test score is based upon "engineering college freshmen." Similarly the student taking both Latin and the required course in English is usually competing with rather select students in the first class. If he has a percentile rank of 50 (exactly average) in Latin and 75 in English, this does not necessarily mean that his knowledge of English is superior to that of Latin. The typical student in the Latin class may have a percentile rank of 75 in English when compared with all students in the English class. This problem does not arise with a battery of tests that have all been standardized upon the same group. It is minimized when some procedure for obtaining equivalent scores is available.

2 Since differences between scores are less reliable than the separate scores themselves, only large discrepancies in the individual's profile should be interpreted as indicating better achievement in one area than in the other. Many teachers try to use for diagnostic purposes slight or moderate differences that may be due entirely to chance. With a group profile based upon average scores in a school grade of even as many as 30 students, however, the differences do not need to be as large in order to be both statistically and educationally significant for averages are much more reliable than the scores from which they are obtained.

3 Percentile ranks do not form an equal unit scale, so they should be spread out at both ends and condensed in the middle when being used as the basis for points in a profile. See Figure 37 on page 297.

²² A master's thesis summarized in Charles C. Peters (Editor), *Abstracts of Studies in Education at The Pennsylvania State College Part IV* 1936, pages 27-28.

²³ Arthur C. Tracker, *The Use of Test Results in Secondary Schools*, *Educational Records Bulletin* 25: 8-9, 1938.

Centile	Theoretical		Economic		Aesthetic		Social		Political		Religious		Centile
	M	W	M	W	M	W	M	W	M	W	M	W	
99	60-62	53-55	62-65	56-58	59-62	62-64	51-56	57-59	52-60	52-53	60-63	67-70	99
97	59	51-52	60-61	51-55	57-58	60-61	52-53	55-56	57	51	58-59	61-66	98
96	58	50	59	53	55-56	58-59	51	54	56	50	56-57	63	97
95	57	49	58	52	54	57	50		55	49	55	62	96
94	56	48	57	51	53	56	49	52-53	54	48	53-54	60-61	94-95
93	55	47	56	50	52	55	48	51	53	47	52	59	93
92	54	46	55	49	51	54	47	50	52	46	51	58	92
91	53	45	54	48	50	53	46	49	51	45	50-51	56	90-91
90	52	44	53	47	49	52	45	48	50	44	49	55	88-89
89	51	43	52	46	48	51	44	47	49	43	48	54	86-87
88	50	42	51	45	47	50	43	46	48	42	47	53	83-85
87	49	41	50	44	46	49	42	45	47	41	46	52	
86	48	40	49	43	45	48	41	44	46	40	45	51	
85	47	39	48	42	44	47	40	43	45	39	44	50	
84	46	38	47	41	43	46	39	42	44	38	43	49	
83	45	37	46	40	42	45	38	41	43	37	42	48	
82	44	36	45	39	41	44	37	40	42	36	41	47	
81	43	35	44	38	40	43	36	39	41	35	40	46	
80	42	34	43	37	39	42	35	38	40	34	39	45	
79	41	33	42	36	38	41	34	37	39	33	38	44	
78	40	32	41	35	37	40	33	36	38	32	37	43	
77	39	31	40	34	36	39	32	35	37	31	36	42	
76	38	30	39	33	35	38	31	34	36	30	35	41	
75	37	29	38	32	34	37	30	33	35	29	34	40	
74	36	28	37	31	33	36	29	32	34	28	33	39	
73	35	27	36	30	32	35	28	31	33	27	32	38	
72	34	26	35	29	31	34	27	30	32	26	31	37	
71	33	25	34	28	30	33	26	29	31	25	30	36	
70	32	24	33	27	29	32	25	28	30	24	29	35	
69	31	23	32	26	28	31	24	27	29	23	28	34	
68	30	22	31	25	27	30	23	26	28	22	27	33	
67	29	21	30	24	26	29	22	25	27	21	26	32	
66	28	20	29	23	25	28	21	24	26	20	25	31	
65	27	19	28	22	24	27	20	23	25	19	24	30	
64	26	18	27	21	23	26	19	22	24	18	23	29	
63	25	17	26	20	22	25	18	21	23	17	22	28	
62	24	16	25	19	21	24	17	20	22	16	21	27	
61	23	15	24	18	20	23	16	19	21	15	20	26	
60	22	14	23	17	19	22	15	18	20	14	19	25	
59	21	13	22	16	18	21	14	17	19	13	18	24	
58	20	12	21	15	17	20	13	16	18	12	17	23	
57	19	11	20	14	16	19	12	15	17	11	16	22	
56	18	10	19	13	15	18	11	14	16	10	15	21	
55	17	9	18	12	14	17	10	13	15	9	14	20	
54	16	8	17	11	13	16	9	12	14	8	13	19	
53	15	7	16	10	12	15	8	11	13	7	12	18	
52	14	6	15	9	11	14	7	10	12	6	11	17	
51	13	5	14	8	10	13	6	9	11	5	10	16	
50	12	4	13	7	9	12	5	8	10	4	9	15	
49	11	3	12	6	8	11	4	7	9	3	8	14	
48	10	2	11	5	7	10	3	6	8	2	7	13	
47	9	1	10	4	6	9	2	5	7	1	6	12	
46	8		9	3	5	8	1	4	6		5	11	
45	7		8	2	4	7		3	5		4	10	
44	6		7	1	3	6		2	4		3	9	
43	5		6		2	5		1	3		2	8	
42	4		5		1	4			2		1	7	
41	3		4			3			1			6	
40	2		3			2						5	
39	1		2			1						4	
38			1									3	
37												2	
36												1	

Figure 11. Centile Sheet for College Men and Women in Allport, Lindzey, and Vernon, *A Study of Values*, Based upon the Norms (851 Men, 965 Women) in the Manual of Directions (Boston: Houghton Mifflin Company, 1951). The Six Scores of John Doe are Plotted

Figure 11 illustrates the above three precautions and also shows how sex differences can be taken into account.²⁴ "Study of Values" scores are shown for a certain college sophomore who secured 67 points on the theoretical scale, which is above the 99th (per)centile, and only 20 points for the economic value. Clearly, this man's dominant value among the six is theoretical. Whether he actually has less of the economic attitude than of the political cannot be decided from this profile, since these two centiles (1 and 3) are negligibly different. Likewise, it is hazardous to say that he is less religious than aesthetic, or less aesthetic than social, for these three

²⁴ An explanation of the basis for this table is contained in Julian C. Stanley, "Study of Values Profiles Adjusted for Sex and Variability Differences," *Journal of Applied Psychology*, 37: 472-473, December, 1953.

values all occur rather close together near the middle of the profile. Therefore, even with such an extreme profile we can safely say only that he is highest on theoretical and social, lowest on economic and political, and not particularly high or low on religious and aesthetic.

Reports to parents or public. Only a few schools make a systematic effort to keep the public informed regarding the educational progress of its schools. Results of the testing program might very well be summarized before the Parent-Teacher Association, women's clubs, luncheon clubs, and similar organizations. Slides and charts, illustrating the nature of the tests, with analysis and interpretation of the records of typical pupils, would be instructive. The cumulative record cards are naturally of great value in conferences with parents regarding the educational program of their children. Hilbert²² points out clearly how this may be done. Unless parents, as well as teachers, administrators, and students, participate in the interpretation of test results and the planning of action based at least partly upon them, the testing program will not be optimally effective. A further discussion of the use of measurement in programs of public relations is given in Chapter 16.

SELECTED REFERENCES FOR FURTHER READING

- Bennett, George K., Seashore, Harold G., and Wesman, Alexander G., *A Manual for the Differential Aptitude Tests* (Second Edition) New York: The Psychological Corporation, 1932. 77 pages.
- Buros, Oscar K. (Editor), *The Fourth Mental Measurements Yearbook*. Highland Park, New Jersey: Gryphon Press, 1953. 1163 pages.
- Buros, Oscar K. (Editor), *Succeeding Mental Measurements Yearbooks* published by the author, Highland Park, New Jersey.
- Coleman, William, *Test Results for Curriculum Study. Annual Report of the Tennessee State Testing Program, 1950-51*. Nashville, Tennessee: Tennessee State Department of Education, 1951. 18 pages.
- Coleman, William, and Cobb, E. B., *The Guidance Use of Test Results*. Knoxville, Tennessee: The Tennessee State Testing Program, University of Tennessee. November, 1951. 47 pages.
- Cronbach, Lee J., *Essentials of Psychological Testing*. New York: Harper & Brothers, 1949. Chapters 4 and 5, "How to Choose Tests" and "How to Give Tests."
- Flanagan, John C., Adkins, Dorothy, and Cadwell, Dorothy H. B., *Major Developments in Examining Methods*. Chicago: Civil Service Assembly, 1313 East 60th Street, November, 1950. 24 pages.
- Jordan, A. M., *Measurement in Education*. New York: McGraw Hill Book Company, 1953. Chapter 4, "The Testing Program—Achievement-Test Batteries."
- Lindquist, E. F. (Editor), *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapters 6, 10, 11, and 12. "Planning the Objective Test," by K. W. Vaughan, "Administering and Scoring the Objective Test,"

²² Robert N. Hilbert, "Parents and Cumulative Records," *Educational Record* 21:172-183. Supplement No. 13. January 1940.

- by Arthur E Traxler, "Reproducing the 'test,'" by Geraldine Spaulding, and "Performance Tests of Educational Achievement," by David G Ryans and Norman Frederiksen
- National Committee on Cumulative Records, *Handbook of Cumulative Records* Washington, D C U S Office of Education, 1944 104 pages
- Stephenson, William, *Testing School Children, An Essay in Educational and Social Psychology* New York Longmans, Green and Company, 1949 Chapter IX, "Principles and Practice of Selection "
- Super, Donald E *Appraising Vocational Fitness by Means of Psychological Tests* New York Harper, 1949 Chapters IV and V, "The Nature of Aptitudes and Aptitude Tests" and "Test Administration and Scoring "
- Traxler, Arthur E , Jacobs, Robert, Selover, Margaret, and Townsend, Agatha, *Introduction to Testing and the Use of Test Results in Public Schools* New York Harper & Brothers, 1953 113 pages
- Worcester, D A , "A Misuse of Group Tests of Intelligence in the School," *Educational and Psychological Measurement*, 7 779-781, Winter, 1947
- World Book Company, *Yonkers-on-Hudson* New York The following free publications, most of them reprints of articles from professional journals, may be of interest in connection with this chapter
- Durost, Walter N , "What Constitutes a Minimal School Testing Program," Test Service Notebook No 1
- Durost, Walter N , "Tests and the Junior High School Guidance Counselor," Test Service Notebook No 2
- Lewin, Lillie, "Pupil Adjustment Through Measurement " Test Service Bulletin No 40
- Burnside, Carolyn J , "Improving the Reading of Seventh Graders," Test Service Bulletin No 44
- Super, Donald E , "The Place of Aptitude Testing in the Public Schools," Test Service Bulletin No 49
- Starkey, Mary L "Determining Individual Needs and Capacities Through Testing," Test Service Bulletin No 56
- Stenquist, John L , "Growth," Test Service Bulletin No 59
- Bridges, Claude F , "Some Basic Considerations in Determining the Significance of Achievement Test Results," Test Service Bulletin No 66
- Brown, Woodrow A , "Testing in Pennsylvania's Public Kindergartens," Test Service Bulletin No 67

9

The Graphical Representation of Educational Data

A. The Value of Graphs

"One picture is worth ten thousand words" So runs an old Chinese proverb "There is a magic in graphs," says a modern writer ¹ He describes the dynamic role of the graphical representation of numerical data as follows

Words have wings, but graphs interpret Graphs are pure quantity stripped of verbal sham, reduced to dimension, vivid, unescapable Wherever there are data to record, inferences to draw, or facts to tell, graphs furnish the unrivaled means whose power we are just beginning to realize and to apply

There can be little doubt that the graphical representation of educational data is a valuable supplement to statistical analysis and summarization The psychological value of graphs in the testing program may be considered under three headings They attract attention, they clarify the meaning, and they aid retention

Graphs attract attention. In the first place, the graph or chart tends to attract the reader's attention Advertisers employ a wide variety of pictures, charts, and diagrams, for they realize that the first step in making a sale is to attract the prospective customer's attention They have learned that pictures will do this where numerical data and printed material will not The average reader is likely to give scant attention to the ordinary printed matter in a school report and be wholly unimpressed by the ap-

¹ Henry D Hubbard quoted by W C Brinton *Graphic Presentation* page 2 New York Brinton Associates 1939

BACK TO SCHOOL

UNITED STATES, 1900 - 1953

ENROLLMENT

MILLIONS

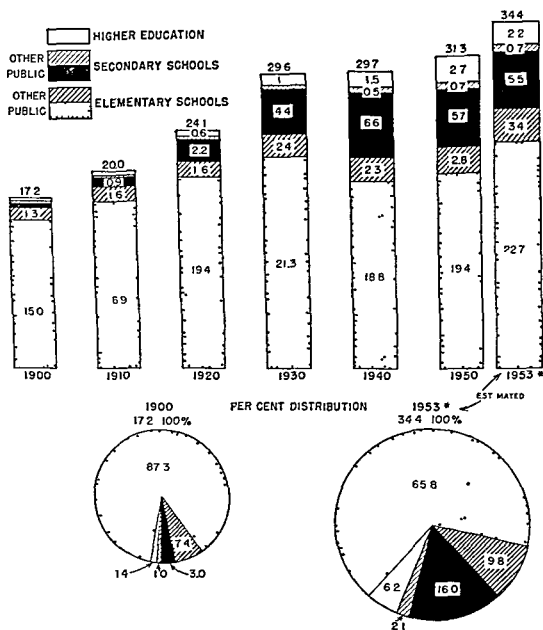


Figure 12 Back to School United States 1900-1953 (Used by Permission of the National Industrial Conference Board)

pulling mass of tabular data often piled up at the end, but his eye is likely to be arrested by any picture or chart that may happen to be included. And this may lead him to read the entire discussion. There is evidence that school administrators are beginning to learn this lesson.¹

Graphs clarify points. In the second place, the graph is often an effective method of clarifying a point. One small chart will often make a point clearer than a dozen tables or paragraphs. It is sometimes said that the facts speak for themselves. In reality, statistics often stand speechless and silent, tables are tongue tied and only the chart cries aloud its message to all the world. Ordinary numerical data are quite abstract; they convey their meaning vaguely and with effort to the average mind. The picture or graph is a more concrete representation of the matter.

Educational facts, such as comparative enrollment figures over a 53 year period, may be presented effectively by graphical means, as shown in Figure 12. There both bar and pie charts are used to contrast the 17.2 million persons enrolled in five types of schools in 1900 with the 34.4 million enrolled in 1953. The bar charts show actual numbers, while the two pie charts have percentage slices.

Figure 13 represents strikingly the enormous inequalities among states in the support of public education.²

A wide variety of charts are shown in Figure 14. The basic information concerning "Motor Buses in Operation in the United States" is given first in tabular form, followed by 15 different black and white charts.³

Graphs aid retention. It has been found that the graphical presentation of certain types of data is a definite aid to recall. Washburne⁴ compared the efficiency of graphical, tabular, and textual modes of presenting historical data to pupils in the junior high school. The material which dealt with certain specific quantitative facts, was kept constant but the mode of presentation varied. Sometimes it appeared as a statistical table, sometimes as a bar graph, a pictograph, or a line graph and at other times it was presented in ordinary paragraph form. Among the conclusions arrived at by the author were the following:⁵

1. The paragraph is in general the form which is least favorable to recall of quantitative data, whether general or specific.

2. The bar graph is the form most favorable to the recall of relative amounts (static comparisons) when the comparisons called for involve a fair degree of diffi-

¹ Cf. Douglas F. Scates, "Reporting, Summarizing and Supplementing Educational Research," *Review of Educational Research* 12: 558-574, December 1942.

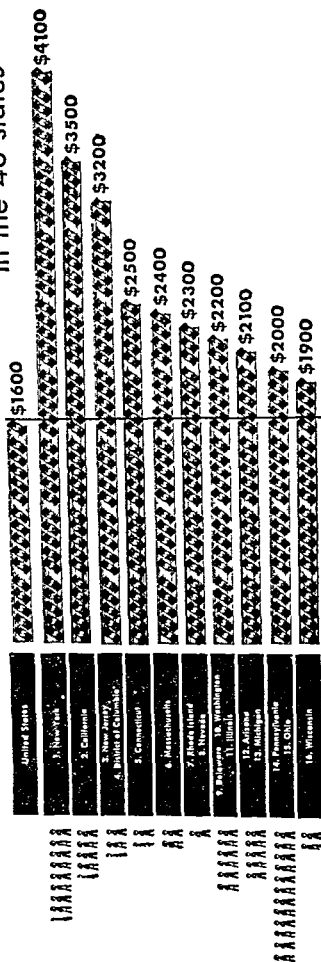
² John K. Norton and Eugene S. Lawler, *Unfinished Business in American Education*, page 13, Washington: American Council on Education, 1946.

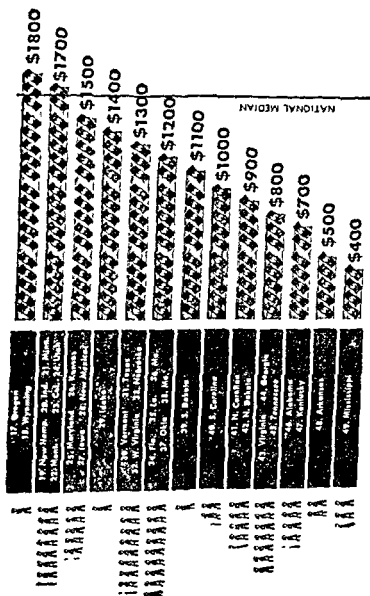
³ This material appears on the first two inside pages of Mary Eleanor Spear's *Charting Statistics*, New York: McGraw-Hill Book Company, Inc., 1952, 253 pages.

⁴ John Noble Washburne, "An Experimental Study of Various Graphic, Tabular and Textual Methods of Presenting Quantitative Material," *Journal of Educational Psychology* 18: 361-476, September and October 1927.

⁵ *Ibid.*, page 475.

Current Expenditure for the Median Classroom Unit in the 48 states





$\text{A} \approx 1\% \text{ of CLASSROOMS in U.S.}$
 $\text{P} \approx \$100$

Figure 13. A Rather Complex Bar Graph with High Attention Value.

WHICH CHART TO USE ? THE DATA

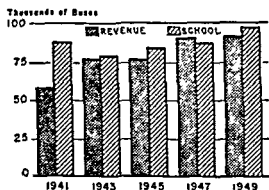
Motor Buses in Operation in United States

Year	Inter-city Buses	Local Buses	Charter and Sightseeing Buses	School Buses	Total Buses
1941	18,420	37,855	2,393	87,400	146,058
1942	22,710	44,101	2,400	79,000	148,211
1943	28,504	45,630	2,000	77,650	153,784
1944	28,000	48,521	3,300	75,500	155,321
1945	29,000	45,955	1,033	83,228	159,216
1946	30,260	47,760	1,475	82,500	161,995
1947	31,900	54,100	3,000	85,900	174,900
1948	3,775	57,175	3,200	90,400	182,550
1949	30,200	57,800	3,500	97,600	189,100

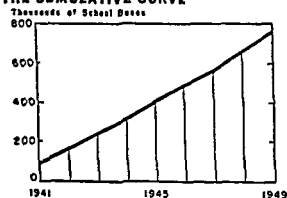
① Data by Inter-city. ② Exchange of students on one bus during school year.

SOURCE: "The Transportation" as of December 31.

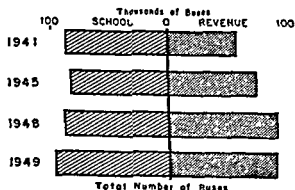
THE GROUPED COLUMN



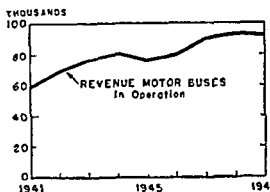
THE CUMULATIVE CURVE



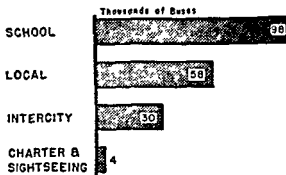
THE SLIDING BAR



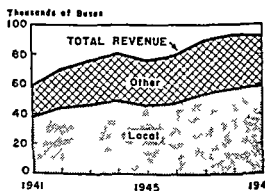
THE CURVE CHART



THE BAR CHART



THE SUBDIVIDED SURFACE



THE PICTOGRAM

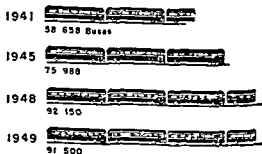
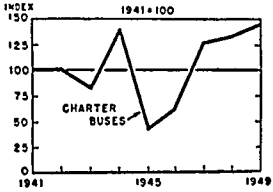
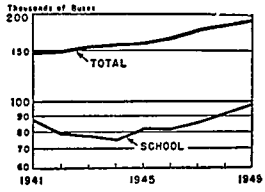


Figure 14 Motor Buses in Operation in the United States—Fifteen Different Charts Based Upon the Same Data

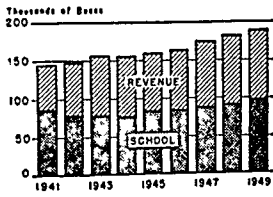
THE INDEX CHART



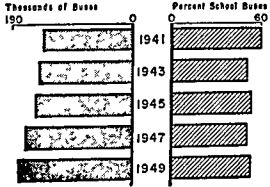
THE LOGARITHMIC CHART



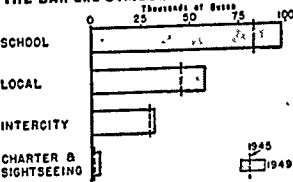
THE SUBDIVIDED COLUMN



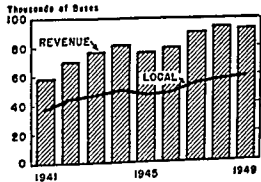
THE PAIRED BAR



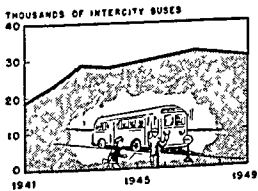
THE BAR and SYMBOL



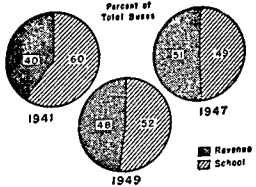
THE COLUMN and CURVE



THE PICTORIAL SURFACE



THE PIE CHART



culty For very simple data some form of pictograph may be more favorable to the recall of relative amounts than the bar graph

3 The line graph is the form most favorable to the recall of relative increase decrease, and fluctuation (dynamic comparisons)

4 The statistical table is the form most favorable to the recall of specific amounts

One study⁷ on graph interpretation in the elementary schools points out that little is known regarding the comparative value of various graphs, although the circle graph appears to be easiest and the line graph most difficult, with the bar graph occupying an intermediate position Her results indicated that a mental age of 14 years was required for the satisfactory interpretation of bar and line graphs without specific instruction in reading materials presented in graphical form

These findings appear to be in line with a principle of learning abundantly supported by experimental evidence namely, that the method of presentation which makes the meaning clearest is most favorable to learning and recall It is important to recognize that neither statistical nor graphical methods bestow precision upon data They are merely useful ways of expressing whatever accuracy exists

B. Representing the Record of an Individual

There is no more striking way of representing the test record of an individual pupil than by means of a graph Such a graphical picture of the strong and weak points of a single person is called a *profile* Sometimes the term *psychograph* is used Many publishers of standard tests provide blank forms for showing these profiles Usually they appear on the first page of the test, where they can easily be detached for filing

Profiles of a single subject. Figure 15 shows the profile of a sixth-grade pupil on the Iowa Silent Reading Tests, New Edition The broken-line profile for the class, based on medians, is also shown With a median standard score of 150, corresponding to a percentile rank of 49 for the eighth month of the sixth grade, John is an average reader The median score of his class is 151, which has a percentile rank of 52 John scored highest on Tests 4 (Paragraph Comprehension) and 6A (Alphabetizing), poorest on Test 6B (Use of Index) He exceeded the class average considerably on Tests 1R (Rate) and 6A (Alphabetizing)

Profiles for a series of subjects. Profiles are especially useful in representing a pupil's record on two or more subjects Most test batteries provide a convenient form for such a profile Figure 16 shows the profile for a tenth grade pupil on the California Achievement Tests, Advanced Battery

This student scored highest on syntax (Grade Placement = 15) and lowest on spelling (GP = 67) His best area is Test 5, Mechanics of English

⁷ Sister Clara Francis Bamberger, 'Interpretation of Graphs at the Elementary School Level' *Educational Research Monographs*, 13 1-62, May 1, 1942

IOWA SILENT READING TESTS

NEW EDITION

By H A GREENE
Director Bureau of Educational Research and Service University of Iowa
and V H KELLEY
University Appointment Office University of Arizona, Tucson, Arizona

Median Score	150
Grade Percentile	49
Grade Equivalent	6.8
Age Equivalent	12-0

Elem.
A
(Revised)
New Edition

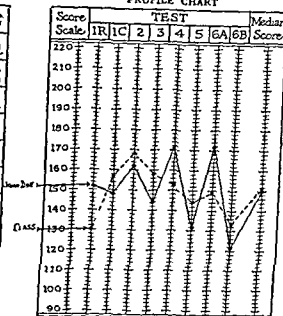
ELEMENTARY TEST: FORM A₁

(Revised)

Name **DOE, JOHN**Age **11** Years **11** Months **6** Grade **6**Set **✓** See On DateMAY 3, 1954 Teacher **MISS SMITH**School **CENTRAL**City and state **METROPOLIS, WIS.**

No.	Test	Score
1	Rate - A + B	152
	Comprehension - A + B	148
2	Directed Reading	162
3	Word Meaning	144
4	Paragraph Comprehension	171
5	Sentence Meaning	131
6	Location of Information	
	A. Alphabetizing	172
	B. Use of Index	121

PROFILE CHART



— JOHN DOE
- - - CLASS

Published 1943 by World Book Company, Yonkers-on-Hudson, New York, and Chicago, Illinois.
Copyright 1933, 1939 by World Book Company. Copyright in Great Britain. All rights reserved. Printed in U.S.A. 1943. 100,000 copies.

NOTE: This test is copyrighted. The reproduction of any part of it by mimeograph, hectograph, or in any other way, whether the reproductions are sold or are furnished free for use, is a violation of the copyright law. Section 8.

Figure 15 Profile of a Pupil and the Sixth Grade Class of Which He is a Member
(Reproduced by Permission of World Book Company)

and Grammar (GP = 12.5). On the complete battery he has a dead-center-average percentile rank of 50 and a grade placement of 10.6.

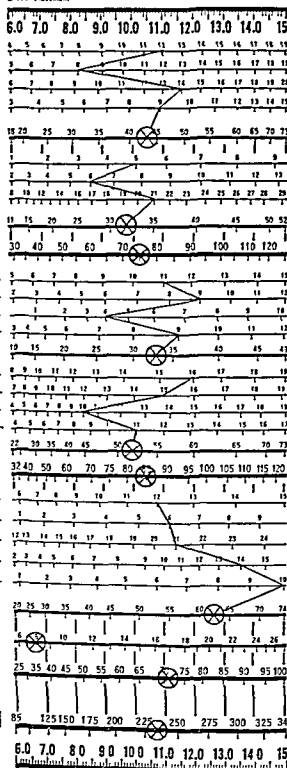
After a period of remedial instruction, the purpose of which is to strengthen the weak points of individual pupils, it is a good practice to give a second form of the same test. A second profile drawn in a different color upon the same sheet is one of the best ways of revealing the progress made, if changes are interpreted cautiously.

SAMPLE PROFILE—COMPLETE BATTERY

A Test Given in January to a 10th Grade Student. Age, 183 Months. Mental Age, 192 Months.

DIAGNOSTIC PROFILE (Chart Student's Scores Here)

Grade Placement



TEST	SECTION	POSSIBLE SCORE	STUDENT'S SCORE
1. READING VOCABULARY	A. Mathematics	22	12
	B. Science	23	8
	C. Social Science	22	14
	D. General	23	9
	TOTAL (A+B+C+D)	90	43
2. READING COMPREHENSION	E. Following Directions	10	5
	F. Reference Skills	15	6
	G. Interpretations	30	21
	TOTAL (E+F+G)	55	32
TOTAL READING		145	75
3. MATHEMATICS FUNDAMENTALS	A. Number Concept	20	11
	B. Symbols and Rules	15	9
	C. Numbers & Equations	10	4
	D. Problems	15	9
	TOTAL (A+B+C+D)	60	33
	E. Addition	20	16
	F. Subtraction	20	15
	G. Multiplication	20	10
4. MATHEMATICS REASONING	H. Division	20	11
	TOTAL (E+F+G+H)	80	52
TOTAL MATH.		140	85
5. MECH. OF ENGLISH AND GRAMMAR	A. Capitalization	15	12
	B. Punctuation	10	6
	C. Words and Sentences	25	21
	D. Parts of Speech	17	14
	E. Syntax	13	10
	TOTAL (A+B+C+D+E)	80	63
6. SPELLING	TOTAL SPELLING	30	8
	TOTAL LANGUAGE	110	71
Handwriting		100	100
TOTAL TEST		395	231
Grade Placement		10.0	
Percentile Rank		90	

INTELL. G.P. 10.3
ACTUAL G.P. 10.4
CHRON. G.P. 9.9

Figure 16. The Profile of a Tenth-Grade Pupil on the California Achievement Test. (Reproduced by permission of the California Test Bureau)

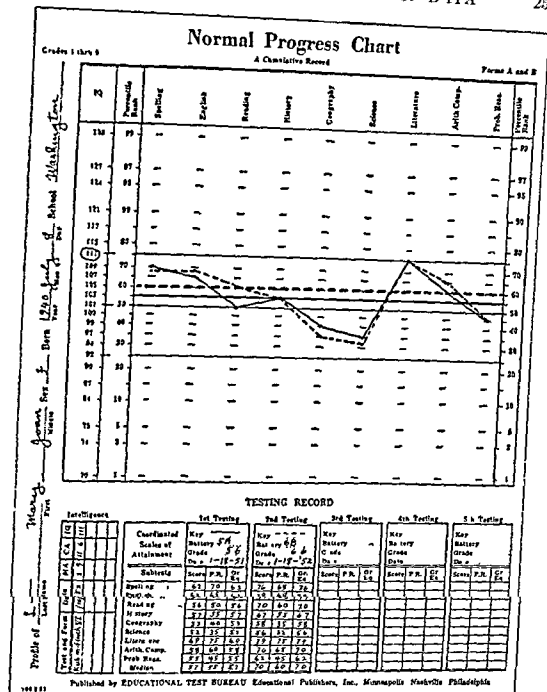


Figure 17. Profiles for a Student Tested in the Fifth and Sixth Grades (Reproduced by permission of Educational Test Bureau)

Figure 17 is a "Normal Progress Chart" issued by the Educational Test Bureau for use with their Coordinated Scales of Attainment. It shows two profiles for Mary L., one (solid line) based upon her performance on Battery 5, Form A, on January 18 in the fifth grade and the other (broken line) for Battery 6, Form B, administered at the same point in the sixth grade. The grade equivalents of her scores varied from 5.2 for science in the fifth grade to 7.9 for English and literature in the sixth grade. Mary's

overall percentile ranks in the fifth and sixth grades are 55 and 60, respectively, while the percentile rank equivalent of her Kuhlmann-Finch IQ of 111 is 75. Thus, even though she is slightly above the average of the class in achievement, Mary seems to be working somewhat below her potential ability to attain

C. Representing a Frequency Distribution

The ordinary frequency distribution does not give a very clear picture of the situation. There are three common methods of representing a distribution of scores graphically: the *histogram* or *column diagram*, the *frequency polygon*, and the *smooth curve*.

The histogram or column diagram. The *histogram* is a series of columns, each of which has as its base one class interval and as its height the number of cases, or frequency, in that class. Figure 18 represents a histo-

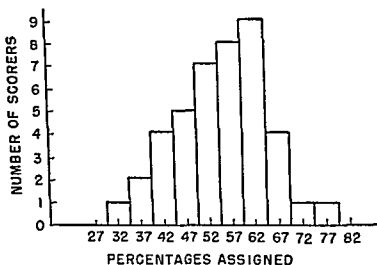


Figure 18 A Histogram, or Column Diagram, Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers

gram showing the distribution of percentage values assigned to an arithmetic paper by forty-two scorers. As the greatest frequency is 9, in the 59.5-64.5 class, it is not necessary to extend the vertical or frequency scale at the left above 9. As the scores range from the 29.5-34.5 class to the 74.5-79.5 class, it is necessary to represent the horizontal scale only through that distance. It is customary, however, to extend the scale one class interval above and below that range. In order to avoid having the figure too flat or too steep, it is usually well to arrange the scales so that the width of the figure is about one and two thirds times its height—that is, the ratio of height to width should be approximately 3/5. In actual practice it is customary to represent the histogram in outline form, rather than to show the full length of the columns. Figure 19 illustrates the shaded outline form of the histogram.

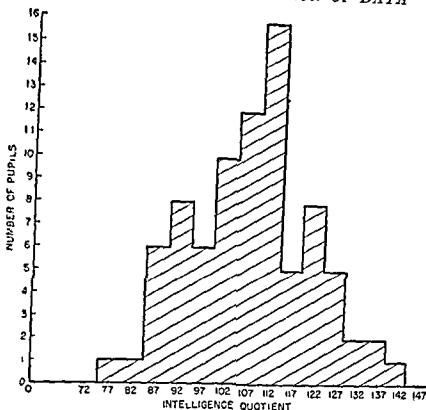


Figure 19. A Histogram, or Column Diagram, Representing the Distribution of the 83 IQs in a Small Junior High School

The frequency polygon. The process of constructing the *frequency polygon* is very much like that of constructing the histogram. In the histogram, the top of each column is indicated by a horizontal line the length of one class interval, placed at the proper height to represent the frequency at that class. But in the polygon a point is located above the *mid-point* of each class interval and at the proper height to represent the frequency

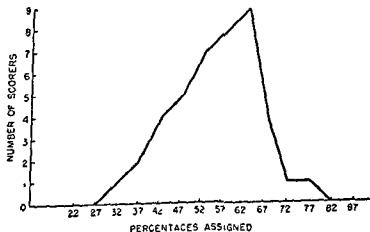


Figure 20. A Frequency Polygon Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers

at that class. These points are then joined by straight lines. As the frequency is zero at the classes above and below those in the distribution, the polygon is completed by connecting the points that represent the highest and lowest classes with the base line at the mid-points of the class intervals next above and below. Figure 20 shows a polygon for the same data represented by a histogram in Figure 18.

The smooth curve. Sometimes a *smooth curve* is drawn instead of the histogram or frequency polygon. The only difference is that for the former a smooth curve is drawn through the points, and for the latter two figures a jagged line is used. The most common use in educational measurement of a smooth curve is in the so-called *normal curve*. Figure 21 shows such a curve superimposed upon a histogram representing the actual distribution of ninth-grade pupils on eleven intelligence tests.

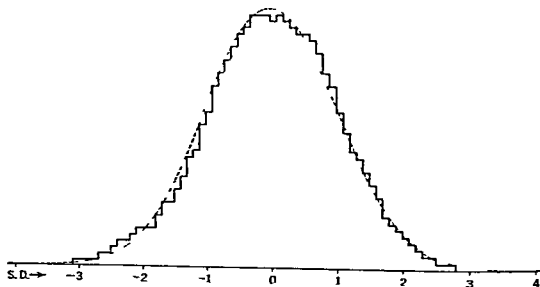


Figure 21 An Actual Curve Compared with the Theoretical Curve of Probability. Actual curve is a histogram based upon scores for eleven well known group intelligence tests administered to the ninth grade. The dotted line represents the theoretical (normal) curve. (From E. L. Thorndike's *The Measurement of Intelligence*, Bureau of Publications, Teachers College, Columbia University, page 529.)

There is one smooth curve, however, which is widely used in representing test scores. This is the *percentile curve*, or *ogive*. Figure 22 shows a percentile curve used to represent the percentage data already employed to illustrate the histogram and the polygon. The points that determine the percentile curve are located on the horizontal line at the upper limit of each class, at the position that indicates on the horizontal scale the percentage of scores up to and including that class. It will be noted, also, that two columns have been added to the ordinary frequency table. The cumulative frequency column indicates the number of scores up to and including each class. For example, there is one score in the 30-34 class, and there are two in the 35-39 class, making a cumulative frequency of 3 in the two lowest

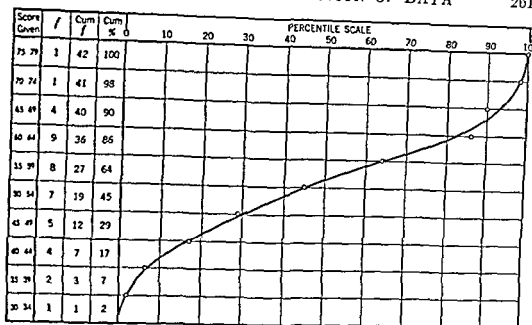


Figure 22. A Percentile Curve Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers

classes. The cumulative per cent column shows what percentage each of these cumulative frequencies is of the total. In the illustration the total, N , is 42. The first entry in this column is, of course, 100, the second is 98, because 41 is 98 per cent of 42; the third is 95, because 40 is 95 per cent of 42; and so on for the others. Each value in the cumulative per cent column is represented as a point on the upper limit of that class interval (the

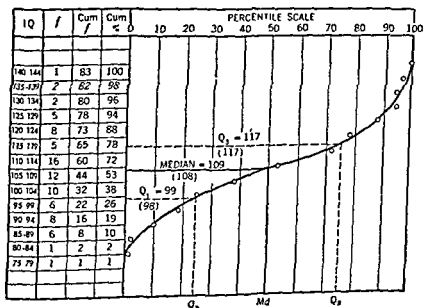


Figure 23. A Percentile Curve Representing the Distribution of 83 IQ's in a Small Junior High School (see Figure 19). The Values of Q_1 , Median (Q_2), and Q_3 Read from the Curve Are Shown with the Computed Values (in Parentheses)

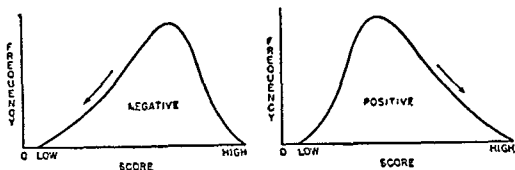


Figure 24 Negative and Positive Skewness

horizontal line separating that class from the class above it), since it includes the percentage of scores up through that class. These points determine the curve. As a rule, especially in small groups where irregularities are most likely to occur, it is best to miss some of the points in order to obtain a smooth and regular curve, but care should be exercised in order to leave about as many points on one side of the line as on the other. Figure 23 shows another ogive. Such a smoothed curve, although it does not exactly represent the actual sampling, probably indicates very closely what is to be expected "in the long run."

Symmetrical and skewed curves. Regardless of whether a distribution is represented as a histogram, a polygon, or a smooth curve, the curve will be either symmetrical in shape, or else pushed or pulled to the right or left. A symmetrical curve is balanced in the center and slopes regularly in both directions. One that is pushed or pulled in one direction is said to be *skewed*. If the peak of the curve is toward the upper end of the scale, with the longest slope downward toward the lower end of the scale, the curve is negatively skewed (skewed to the left). On the other hand, if the peak of the curve is toward the lower end of the scale, with the longest slope toward the higher end of the scale, the curve is positively skewed, or skewed to the right. Both kinds of curves are shown in Figure 24. Many

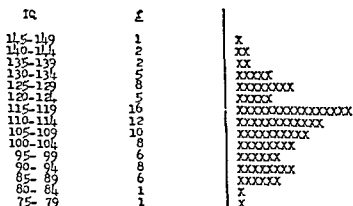


Figure 25 Bar Graph Made on the Typewriter, Showing the Distribution of 91 IQs in a Junior High School

curves met with in educational measurement show some skewness, although the departure from symmetry is usually not very great in larger samplings unless some selective factors are operating

Typewriter graphs. A satisfactory *bar graph* can be made on the typewriter. Figures 25, 26, and 27 illustrate this type of graph. Other graphs, such as the *circle*, or *pie graph*, and various *picture graphs* or *pictographs* are occasionally met with in educational measurement. These are illustrated in Figures 12, 13, and 14.

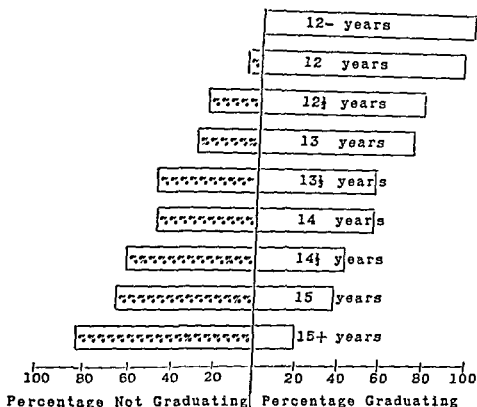


Figure 26 Bar Graph Made on the Typewriter Showing the Percentage of Pupils of Each Age Group Who Were Graduated from High School and the Percentage Who Entered High School but Did Not Graduate

Which graph is best? As is to be expected, no one type of graph is equally good for all purposes. The histogram is the easiest of all to understand and is usually best if but one distribution is being represented. If two or more distributions are to be compared, however, polygons are usually better, since so many lines coincide when histograms are superimposed that the picture is likely to be confusing. The percentile curve has many advantages not possessed by other curves. The first of these is that it is possible to estimate with a high degree of accuracy the quartiles, medians, and other similar points. This means that one can read directly from the curve percentile measures like those illustrated in Figure 23. As will be shown in the next section, by means of percentile curves several groups

can be presented, for convenient comparison, on a single sheet. The principal value of bar graphs, circle graphs, and picture graphs lies probably in school publicity and in the motivation of learning. "A successful graph," as Scates points out, "depends far more on careful thought and judgment than on techniques."⁸

D. Representing Two or More Distributions

There are many occasions when it is desirable to compare two or more distributions. For example, school administrators may wish to compare the intelligence or achievement of the pupils in various classrooms or buildings. The overlapping among the various grades within a single building is a striking way to present the need for individualized instruction and varied materials.

Representing entire distributions. When it is important to compare two or more entire distributions, as would be the case in a study of the status of a school or school system, the choice will usually lie between the frequency polygon and the percentile curve. The difficulty of superimposing two or more histograms has already been pointed out. A series of polygons may be drawn on the same sheet one above the other, or alongside each

Score	Frequency for Grade			Bar Graph for School Grade		
	7	8	9	Seventh	Eighth	Ninth
200-219			3			999
180-199	1	4	5	7	8888	99999
160-179	3	3	7	777	888	9999999
140-159	4	9	7	7777	888888888	9999999
120-139	11	7	11	77777777777	8888888	99999999999
100-119	4	7	2	7777	8888888	99
80-99	4	2	1	7777	88	9
60-79	1	3		7	888	
40-59		1			8	
20-39		1			8	

Figure 27 Graph Made on the Typewriter Showing the Overlapping of Grades Seven, Eight, and Nine in Reading Comprehension

other. Figure 27 illustrates a method of showing overlapping by bar graphs made on the typewriter.

The use of polygons. The distinct advantage of polygons over histograms for representing a series of distributions is that polygons can be superimposed upon each other with less crossing of lines. In this form com-

⁸ Douglas C. Scates *op cit* page 568

parisons among distributions are more easily made. Figure 28 illustrates this possibility with the distribution of reading comprehension scores on the Iowa Silent Reading Test for the seventh, eighth, and ninth grades of a certain school. One fact stands out clearly, the great overlapping of the

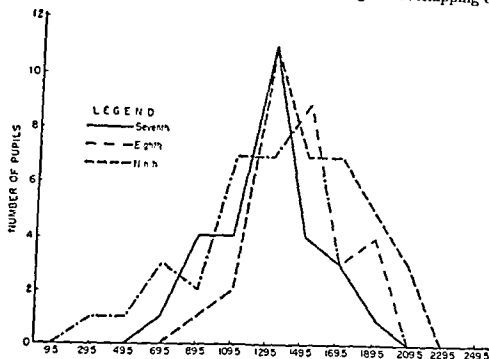


Figure 28 Frequency Polygons Representing the Distribution of Reading Comprehension Scores on the Iowa Silent Reading Tests for the Seventh, Eighth, and Ninth Grades of a Certain School (Data from Figure 27)

three grades in reading ability. But even with only three distributions the lines cross and recross so many times as to make any accurate comparison of one grade with another somewhat difficult. More than three classes can hardly be represented in the same graph by frequency polygons without considerable confusion. It is also difficult to compare distributions accurately where the numbers of cases vary greatly, unless each frequency is represented as a per cent of its total.

The use of percentile curves. For the graphic comparison of two or more distributions the percentile curve has certain outstanding advantages. Since the frequencies are reduced to per cents, it is readily possible to compare groups of unequal size. Another important advantage is that several distributions can be represented in a single graph without difficulty or confusion. Figure 29 shows the distribution of reading comprehension scores for the same grades as in Figure 28 in the form of a percentile curve.

From these percentile curves several relationships are observable that were not apparent in the polygons. It is quite clear that although the seventh and eighth grades have almost exactly the same average scores, the eighth grade has greater variability. This is evident from the fact that

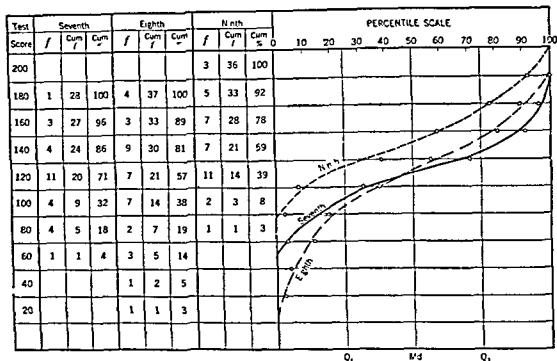


Figure 29. Total Comprehension Scores on the Iowa Silent Reading Tests for the Seventh, Eighth, and Ninth Grades

the upper half of the eighth grade exceeds the upper half of the seventh grade, but that the lower half of the eighth grade falls behind the lower half of the seventh

Furthermore, although the ninth grade runs rather consistently above the other two grades, about 15 per cent of the ninth grade pupils fall below the median of the seventh and eighth grades

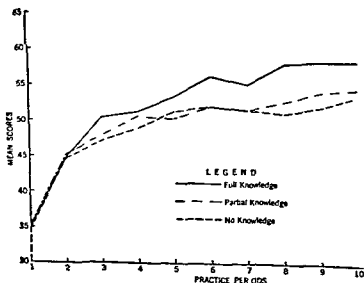


Figure 30 The Learning of Three Groups Compared One with Full Knowledge of Progress One with Partial Knowledge of Progress, and One with No Knowledge of Progress

Representing central tendencies of a series of distributions. It is frequently necessary to represent, not the entire distribution, but only the central tendencies or averages. A learning or progress curve is an illustration. Figure 30 shows a graphic picture of the results of a learning experiment. It shows three groups, one with no knowledge of progress, one with partial knowledge of progress, and one with full knowledge of progress. It will be noted that after the second trial the progress was roughly proportional to the amount of knowledge possessed. A simple line graph makes this clear.

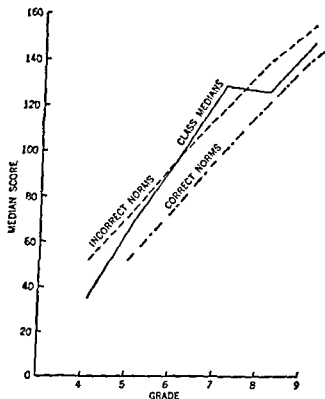
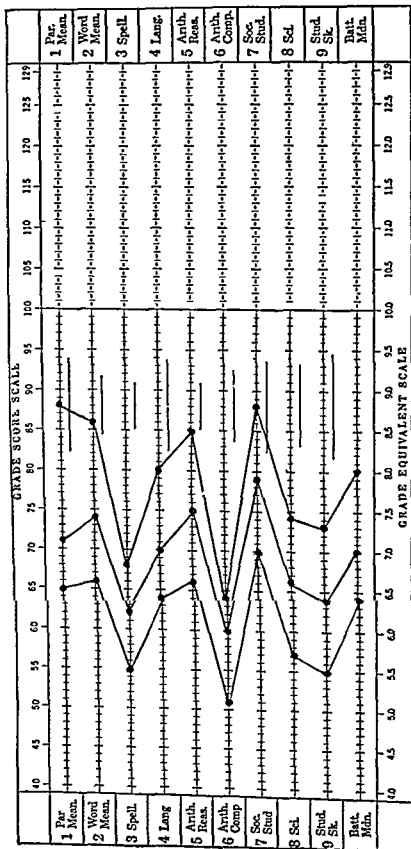


Figure 31. Correct and Incorrect Location of the Norms in a Line Chart Showing Median Scores on a Reading Test

Another common use of the line graph is for comparing two or more schools through several grades, or one school with the norms on a test. Figure 31 shows the correct and the incorrect construction of such a graph. The solid line connects the median scores on a reading test for grades four to nine, inclusive. The tests were given in October, or one-tenth of the way through the grade. The dash line connects the norms *incorrectly* drawn from norms in the manual for the *end of the grade*. The dot dash line connects the norms at the proper grade location. It will be noted that when the line is incorrectly located only the seventh grade appears to exceed the norm, whereas in reality every grade does. Here the horizontal axis is considered a scale and the points determining the lines are located with reference to it.



Grade equivalent values above 10.0 are extrapolated values and not to be interpreted as signifying the typical performance of pupils of the indicated grade placement. (See Directions for Administering.)

Figure 32. Grade Profiles for the Seventh, Eighth, and Ninth Grades of a Certain Junior High School Made by Connecting the Median Scores on Each Part of the Stanford Achievement Test, Advanced Battery Complete, Form J. (Reproduced by permission of World Book Company)

Figure 32 shows the profiles for the seventh, eighth and ninth grades of a certain junior high school made by connecting the median scores on each part of the Stanford Achievement Test. This figure shows clearly that the school is weak in spelling, arithmetic, computation, and study skills and particularly strong in the social studies. It is evident that this school is stressing the content subjects at the expense of some of the more formal tool subjects. Whether or not this appears to be a desirable emphasis depends upon one's philosophy of education.

Representing the central tendencies and variabilities of a series of distributions. The variabilities as well as the central tendencies of a series of distributions may be shown in a similar manner by line graphs. Figure 33 is an illustration. This figure shows Q_1 , the median, and Q_3 for

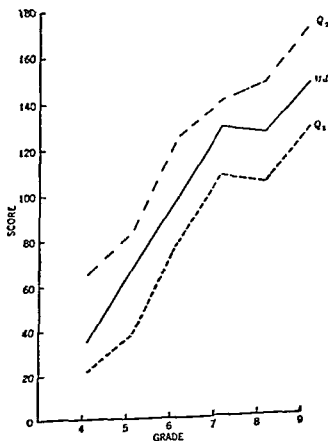


Figure 33 A Line Graph Showing the Medians and Quartiles for Grades Four to Nine Inclusive in Reading Comprehension

each grade from four to nine inclusive in reading comprehension. While the three lines have the same general shape, they converge slightly at the seventh grade where the variability is least. It would be possible to include from the table of norms the corresponding medians and quartiles for the typical school, but to do so would make the figure too complicated for easy interpretation.

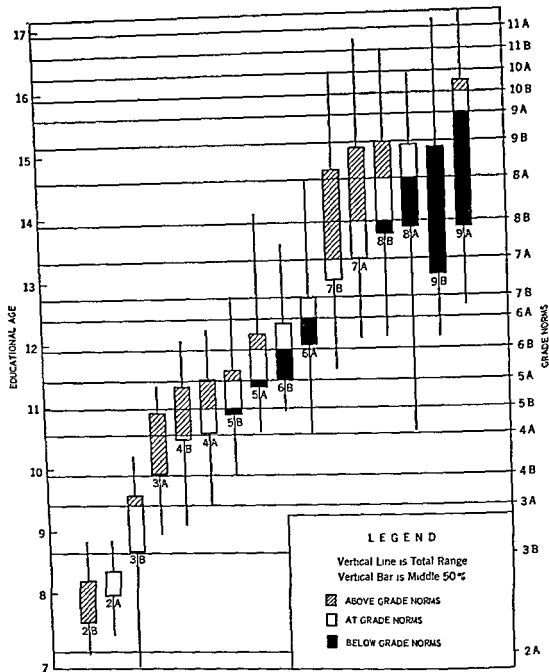


Figure 34 The Central Tendency and Variability in Educational Age of Grades 2B to 9A, Inclusive, in a Small City School System

Figure 34 is a bar graph which shows the central tendency and variability in educational age of grades 2B to 9A, inclusive, in a small city school system.⁹ In each grade the vertical line indicates the total range, the vertical bar indicates the range of the middle 50 per cent, and the middle of the bar is the approximate position of the median. The horizontal lines across the full width of the graph indicate the norms for the beginning of

⁹ Report of the Public Schools of Shelbyville, Kentucky, page 73. Bulletin of the Bureau of School Service, Vol I, No 1. Lexington University of Kentucky, 1928

each grade. It will be noted that the part of each bar which is crosshatched indicates the proportion that is above the grade norm while the shaded part is the proportion that is below the grade norm. The overlapping is especially marked from 7B to 9B. This condition suggests the advisability of trying to find out whether these ninth grade classes happened to be weaker than usual, or whether the teaching emphasis was responsible for the apparent lack of improvement. This type of graph is an effective means for presenting the essential features of a total situation. Here the amount of overlapping is impressive. It will be noted for example that those whose EA is 12-6 are found in all grades from 5B to 9A, and that pupils classified in 8A vary in EA from just above the 4A level to almost the 10A level.

L. General Suggestions for Constructing Graphs

Varied practice. A wide diversity of practice will be found in the construction of graphs as used in psychology and education. The title is sometimes placed above the graph though usually it is placed below. In nearly all books and periodicals the graph title is placed below but in unpublished charts such as wall charts the title is often more effective when lettered above. The figures are numbered consecutively with Arabic numerals placed at the beginning of the title. Sometimes the title is written in capital letters, as in tables, sometimes the initial letters of all important words are capitals, and again, only the first word in the title is capitalized unless there are proper names, in which case the usual rules for capitalization apply. The second of these methods is perhaps most common.

Suggested standards. Years ago a committee composed of representatives of the various groups interested in graphical methods prepared a report¹⁰ recommending certain standards for constructing graphs. This report still covers most of the points required for the proper representation of educational data. The following rules are taken from it.

- 1 The general arrangement of a diagram should proceed from left to right.
- 2 Where possible represent quantities by linear magnitudes as areas or volumes are more likely to be misinterpreted.
- 3 For a curve the vertical scale whenever practicable should be so selected that the zero line will appear on the diagram.
- 4 If the zero line of the vertical scale will not normally appear on the curve diagram the zero line should be shown by the use of a horizontal break in the diagram.
- 5 The zero lines of the scales for a curve should be sharply distinguished from the other coordinate lines.
- 6 For curves having a scale representing percentages it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

¹⁰ W. C. Brinton, Chairman, Preliminary Report, Joint Committee on Standards of Graphic Representation, *Quarterly Publications of the American Statistical Association* 14: 790-797, 1915.

7 When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time

8 When curves are drawn on logarithmic coordinates the limiting lines of the diagram should each be at some power of ten on the logarithmic scales

9 It is advisable not to show any more coordinate lines than necessary to guide the eye in reading the diagram

10 The curve lines of a diagram should be sharply distinguished from the ruling

11 In curves representing a series of observations, it is advisable whenever possible to indicate clearly on the diagram all the curves representing the separate observations

12 The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top

13 Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes

14 It is often desirable to include in the diagram the numerical data or formulae represented

15 If numerical data are not included in the diagram, it is desirable to give the data in tabular form accompanying the diagram

16 All lettering and all figures on a diagram should be placed so as to be easily read from the base or the bottom, or from the right hand edge of the diagram as the bottom

17 The title of a diagram should be made as clear and complete as possible Sub-titles or descriptions should be added if necessary to insure clearness

A useful manual which treats of the different phases of the construction of line charts has been prepared by the Committee on Standards for Graphic Presentation¹¹ For a fuller discussion of the general problem of graphical representation, several excellent books listed at the end of this chapter are available

The suggestions given by Spear should be kept constantly in mind when constructing graphs¹²

In the present day, when visual education in all aspects has become, not only an aid to but also a vital basis of learning, our attention is called more than ever before to the almost limitless possibilities in this field. The eye absorbs written statistics but only slowly does the brain receive the message hidden behind written words and numbers. The correct graph, however, reveals that message briefly and simply. Its purposes, which follow, are clear from its context.

- 1 Better comprehension of data than is possible with textual matter alone
- 2 More penetrating analysis of subject than is possible in written text
- 3 A check of accuracy

This triple purpose of the chart can be carried out through careful planning and familiarity with the functions of all types of graphs and media. The following six steps are fundamental to the development of graphic presentation that will describe statistical data with clarity and dramatic impact.

¹¹ *Time Series Charts: A Manual of Design and Construction*, 68 pages. New York: American Society of Mechanical Engineers, 1938.

¹² Quoted by permission from pages 3-4 of Mary Eleanor Spear, *Charting Statistics*. New York: McGraw-Hill Book Company, Inc., 1952.

- 1 Determine the significant message in the data
- 2 Be familiar with all types of charts and make the correct selection
- 3 Meet the audience on its own level, know and use all appropriate visual aids
- 4 Give detailed and intelligible instructions to the drafting room
- 5 Know the equipment and skills of the drafting room
- 6 Recognize effective results

Even when no technical assistance is available, teachers and administrators can make excellent use of graphs to facilitate the attainment of educational objectives

SELECTED REFERENCES FOR FURTHER READING

- Arkin, Hubert, and Colton, Raymond R, *Graphs How to Make and Use Them* New York Harper & Brothers, 1936 224 pages
- Brinton, Willard Cope, *Graphic Presentation* New York Brinton Associates, 1938 512 pages
- Kelley, Truman L, *Fundamentals of Statistics* Cambridge, Mass Harvard University Press, 1947 Chapter IV, "Graphic Methods"
- Modley, Rudolph, *How to Use Pictorial Statistics* New York Harper & Brothers, 1937 170 pages
- "Presentation Problems," a feature in the *American Statistician* since August 1947
- Spear, Mary Eleanor, *Charting Statistics* New York McGraw-Hill Book Company, 1952 253 pages
- Thompson, Loring M, "Meaning in Space," *ETC A Review of General Semantics*, 8 193-201, Spring, 1951
- Vernon, M D, "The Use and Value of Graphical Methods of Presenting Quantitative Data," *Occupational Psychology, London*, 26 22-34, January, 1952

10

The Uses and Limitations of Norms

It is self-evident that the value of test scores will be dependent largely upon how well they are understood. The preceding chapter concerned the summarization of scores by graphical methods as an aid to their interpretation. The present chapter will consider some closely related problems of interpreting scores by the aid of norms.

A Norms and Standards

Standardized versus nonstandardized tests. At the outset it is important to distinguish clearly between a *norm* and a *standard*,¹ especially because the terms are frequently used interchangeably. The confusion doubtless arises over the fact that norms are used with standard tests and that a part of the process of standardization is the derivation of norms.

Many standard tests began as informal objective tests made by classroom teachers. When an informal test has gone through the process of standardization, it then differs from the original class test in four essential aspects. In the first place the content has been standardized. This means that each item has survived most careful scrutiny by a competent person, or more likely a group, and that its difficulty and value have been determined by rigid experimental processes that have eliminated its weaker fellows. In the second place its method of administration has been standardized. This means that definite directions have been worked out, usually with appropriate time limits and the like. In the third place, the method of scoring has been standardized. This means that scoring keys have been

¹ John C. Flanagan emphasizes this distinction on page 698 and elsewhere in *Units, Scores, and Norms*. Chapter 17 in E. F. Lindquist (Editor), *Educational Measurement*. Washington, D. C.: American Council on Education, 1951.

prepared and that definite rules have been formulated for marking the papers and for determining the scores on each part and on the whole test. Finally, the process of interpretation has been standardized at least in part. This means that tables of norms are now available for interpreting the various scores made on the test. These norms are merely scores which have been made by large numbers of pupils distributed over wide geographical areas and representing various types of schools, and which have been grouped, as a rule, according to chronological age or school grade.

Norms versus standards. The word *standard* implies a *goal* or *objective to be reached*. It should be clear, then, that a *norm* is not a measure of *what ought to be*, a *goal*, but is merely a measure of *what is*, the *status quo*. When a grade or class is up to the national median on the test, it is just an average or typical group. Of course, it may be that this score represents a reasonable performance for the group under the circumstances, but that fact would have to be determined by further inquiry. The mere fact that the grade attains the norm does not of itself establish anything other than that the performance is that of a typical group. Manifestly a group of students having superior opportunities and capacities ought to make better than a typical record. On the contrary, a group of low ability and opportunity might find it virtually impossible to do that well. Unfortunately, at the present time not many tests have more than one set of norms for each grade or age group, all types of pupils and schools being lumped together.² What is needed is a norm for at least each major type of school organization and type of pupil. Even then such norms could hardly be regarded as reasonable standards of attainment. For one thing, the norms of achievement tests are never more than tentative. They must be continually changing with increases in length of school term and with improvement in training of teachers, in textbooks, in school equipment, and the like. It is also not unreasonable to assume, human nature being what it is, that average achievement with the facilities now available could be considerably better than exists at the present time. In a real sense the only valid norm for the individual pupil is his own past record, and the only valid standard is his maximum capacity for growth.

Reasonable standards or goals of attainment, are almost altogether lacking. It is conceivable that such standards might be worked out and expressed in numerical units on existing tests, or on others to be devised. But such a process is inherently difficult, whereas the process of building norms is time-consuming and laborious but perfectly simple and straight forward. In fact, an adequate technique for establishing standards has yet to be worked out. Ideally, a standard would have to be provided for each individual. At any rate, no one standard could be established which would

² The achievement tests prepared by the Armed Forces Institute had separate norms for six geographical regions as well as for the country as a whole. See *Educational Record* 25:369 October 1944.

be equally appropriate for every body, or even for any considerable number. In view of such considerations as these, Wood has said¹

As currently used, the word *standard* has no place in educational literature outside the perorations of convention orators

Speaking more constructively, it is sufficient to point out that educational standards are necessarily individual and in their fundamental nature are akin to the standards of tailors and shoemakers who judge the quality of their products by how well they fit the individual for whom they are intended and how long they serve him

Swan has satirized the idea of a single uniform standard by imagining what would happen if all the tailors of the country got together and agreed upon a "standard suit"⁴ The distressing outcome is described⁵ as follows

Instead of the old haphazard procedure, the standard suit was brought out when a man went into a tailor shop to get a new suit. If he did not fit the suit, he was rejected then and there. He was thus sentenced to join a nudist colony. Men soon learned that the only thing to do was to eat the right food and take the proper exercise to make them just fit the suit. If he perchance ate something else than that required to make him fit the standard suit, he would be rejected, even though what he ate was better for him from the standpoint of health than that needed to get ready for the standard

The important thing to remember is that, for the present at least, such standards do not exist in any subject. Certainly an understanding of the way norms are determined would make it obvious that they lay no claims to being goals of performance, unless perchance one is willing to accept mediocrity as a goal

B Raw Scores and Derived Scores

What a score means. To take a simple case, let us suppose that a certain pupil has made a score of 40 on a spelling test of 50 words. What does this score of 40 mean? To say that the score represents an achievement of 80 per cent is true as far as it goes, but this obvious interpretation leaves much to be desired. As the problem of interpreting a given score in meaningful terms is fundamental in all measurement, it deserves careful consideration

A score on any test is simply a *numerical description of an individual's performance on that test*. A distinction must be made between test performance on the one hand and ability and capacity on the other hand. Performance is merely evidence of ability or capacity. *Ability* refers to an individual's *actual achievement* at the present time, whereas *capacity* refers

¹ Ben D. Wood, *Basic Considerations*, *Review of Educational Research* 3, 13 February, 1933

⁴ J. N. Swan, 'Standardized Tests for Chemistry Teaching', *School and Society* 44, 275-277, August 29, 1936

⁵ *Ibid.* page 276

to his *potentialities*. Since a test is always a sampling rather than a complete measurement a pupil's response to the test situation is accepted as an expression of his ability operating under a given set of conditions. But a poor score on a valid achievement test is not necessarily evidence of poor ability in that subject under any and all conditions. It may be due to any number of factors such as physical illness or discomfort, poor eyesight or hearing, emotional disturbance or dislike for the teacher or subject.

In like manner a poor performance on even the best group test of intelligence available is not necessarily positive proof of a lack of what we call 'general intelligence.' It may be due to any one factor or combination of factors mentioned above as operating in the case of achievement tests. In addition there are several other factors that may be responsible such as poor reading ability, inability to understand the English language and especially inadequate learning opportunities in school and outside. For example, Wheeler⁶ found that the average intelligence of Tennessee mountain children as measured by two well known group tests was approximately normal at six years but that it showed a fairly consistent decrease with increases in chronological ages. The data warrant the significant conclusion

The general trend of this investigation indicates that the results of both tests are materially affected by environmental factors and that the mountain children are not as far below the normal as the tests seem to indicate. With the proper environmental changes the mountain children might test near a normal group.

Ten years later Wheeler⁷ repeated the study in the same region which had shown 'definite improvement in the economic, social and educational status during the intervening period. Although there was still a tendency for intelligence as measured by the tests to decline in the upper years, the average IQ was ten points higher than it had been a decade earlier.

A study of Kentucky mountain children by Asher⁸ revealed similar results and led to the conclusion that a valid comparison of the intelligence of urban children and of children in less favorable environments awaits more adequate measuring methods.

A group of researchers at the University of Chicago found that many intelligence test items were answered correctly much more frequently by children from the higher socio-economic levels than by those from the lower ones and they concluded that the tests were penalizing the latter youngsters for their lack of contact with middle-class culture and lack of appre-

⁶ L. R. Wheeler, 'The Intelligence of East Tennessee Mountain Children,' *Journal of Educational Psychology* 23, 351-370, May 1932.

⁷ L. R. Wheeler, 'A Comparative Study of the Intelligence of East Tennessee Mountain Children,' *Journal of Educational Psychology* 33, 321-334, May 1942.

⁸ E. J. Asher, 'The Inadequacy of Current Intelligence Tests for Testing Kentucky Mountain Children,' *Journal of Genetic Psychology* 46, 480-486, June 1935.

eration for middle class behavior.⁹ This interpretation has stirred up considerable discussion, many psychologists do not accept the Chicago group's conclusion that the items should be rewritten to be "fairer" to lower class persons.¹⁰

The point is that capacity is always inferred from activity or performance. The inference, for example, that two identical scores on an intelligence test really mean equal degrees of intelligence cannot be safely made unless it is known that the learning opportunities have been at least approximately equal. A full realization of this fact would enjoin more caution than is often shown in the interpretation of scores on so-called tests of general intelligence. Trained examiners exercise care in observing rigidly controlled conditions for administering the tests and objective standards for scoring the papers, but it is often hard to be sure about the pupil's past history, which may be reflected to some extent in his present performance.

Raw scores versus derived scores. When a test paper has been marked according to instructions, the score obtained is called a *raw* score or *crude* score. On tests as distinguished from quality scales it is often called a *point* score since the numerical description is in terms of points. On a scale as for example the Ayres handwriting scale, the numerical description is hardly in terms of points but rather in terms of some arbitrary value assigned to a rank or position. In the example given above, the pupil has a point score of 40 on the spelling test. In other words, 40 describes his performance on that particular test at the time it was administered.

But a raw or point score by itself means very little. It is usually not possible to compare a raw score on one test directly with a raw score on another test. The difficulty is that the units are not comparable. The problem is much like that imposed when adding $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$ and $\frac{5}{8}$. It is first necessary to find a common denominator, in this case 12, and then to express all values in terms of that denominator. The problem is then simple:

$$\frac{1}{2} + \frac{2}{3} + \frac{3}{4} + \frac{5}{8} = \frac{6}{12} + \frac{8}{12} + \frac{9}{12} + \frac{7.5}{12} = \frac{30.5}{12} = 2\frac{7}{12}$$

⁹ The most comprehensive report of these studies is Kenneth W. Eells and others, *Intelligence and Cultural Differences*, 388 pages, Chicago: University of Chicago Press, 1951. Interesting background reading is Allison Davis, *Social Class Influences upon Learning: The Inglis Lecture, 1948*, 100 pages, Cambridge, Mass.: Harvard University Press, 1948.

¹⁰ Two essentially unfavorable reviews of Eells' book are John G. Darley, "Review: Intelligence and Cultural Differences," *Journal of Applied Psychology* 36, 141-143, April 1952, with a reply by Eells, "Comment on Darley's Special Review," *Journal of Applied Psychology* 36, 422-423, December 1952; and Quinn McNemar, "Review: Intelligence and Cultural Differences," *Psychological Bulletin* 49, 370-371, July 1952. At the 1952 American Psychological Association convention in Washington, D. C., there was a symposium entitled "Implications of the Chicago Studies of Intelligence and Cultural Differences" (see the *American Psychologist* 7, 290, July 1952). Speakers were Eells, Roger T. Lennon, Irving Lorge, and John L. Stenquist. Also see "Techniques for the Development of Unbiased Tests," pages 76-131 in *Proceedings of the 1952 International Conference on Testing Problems*, Princeton, N. J.: Educational Testing Service, 1953. Papers were presented by Ernest A. Haggard of the University of Chicago, Irving Lorge of Teachers College, Columbia University, and Philip J. Rulon of Harvard University, with a written commentary by McNemar to which Haggard replied.

To meet a similar need, test makers have found it necessary to determine common denominators for their test scores. These are called "derived scores." *A derived score is a numerical description of a pupil's performance in terms of norms.* The norm itself is the performance of a defined group considered to be typical. For example, a pupil who answers correctly 22 questions on the Thorndike-McCall Reading Test has a reading ability which is that of the normal, or average, twelve-year old child at the end of the fifth grade.

Usually, with standard tests the norms used are either age norms or grade norms. The derived scores merely describe the individual's position in some group. Sometimes the age norms are carried one step further and expressed in terms of quotients, that is, one age score is divided by another. With the exception of quotients, most derived scores are obtained from tables of norms in the test manuals which give in parallel columns the derived scores equivalent to various point scores. As the problems of interpretation differ somewhat for achievement and intelligence tests, they will be treated separately in the next two sections.

C. The Use of Norms in Interpreting Scores on Intelligence Tests

Mental age versus intelligence quotient The most commonly used units in which to express the results of an intelligence test are mental age and intelligence quotient, usually abbreviated MA and IQ.¹¹ It is important to understand the distinction between them. MA is a measure of *mental maturity* and "indicates the level of development which a child has reached at a given time," to use the words of Terman.¹² This degree of mental maturity or level of development is expressed in terms of that "*possessed by the average child of corresponding chronological age*."¹³ For example, a point score of 75 on the Terman Group Test of Mental Ability is equivalent to an MA of 13 years and 2 months usually written 13.2. This means that when the Terman Test had been given to hundreds of children in various parts of the country it was found that the average score of a child with a chronological age (CA) of 13 years and 2 months was 75 points. Any child who makes a score of 75 on this test is said to have an MA of 13.2. But pupils of various CA's make scores of 75 on the Terman Test. It is clear, therefore, that a 10-year-old child with an MA of 13.2 has matured rapidly, whereas a 14-year old child with an MA of 13.2 has matured at a much slower rate. In other words, MA is a measure of *stage* or *level* of maturity but *not* of *rate*. Rate is indicated by the IQ which is obtained by dividing the MA by the CA and multiplying the resulting

¹¹ Standard score IQ's which do not involve mental ages are used with the Wechsler Intelligence Scales both adult and child. They possess certain distinct advantages over the traditional IQ and are popular with individual testers.

¹² Lewis M. Terman, *The Intelligence of School Children*, page 7. Boston: Houghton Mifflin Company, 1919.

¹³ *Ibid*.

decimal fraction by 100. $IQ = 100 \left(\frac{MA}{CA} \right)$ In the preceding illustrations the IQ of the 10-year-old child whose MA is 13-2 would be:

$$\begin{aligned} IQ &= 100 \left(\frac{MA}{CA} \right) = 100 \left(\frac{13-2}{10} \right) = 100 \left(\frac{13\frac{2}{5}}{10} \right) = 100 \left(\frac{13\frac{2}{5}}{10} \right) \\ &= \frac{10 \left(\frac{13\frac{2}{5}}{10} \right)}{100} = 131\frac{2}{5} = 132. \end{aligned}$$

The IQ, then, gives us a different interpretation of a score on an intelligence test from that afforded by the MA. The *IQ is a measure of rate of maturity*, whereas the *MA is a measure of level or stage of maturity*. In both cases rate and level are relative to the standardization group. If a child has matured rapidly, he is said to be bright, if he has matured slowly, he is said to be dull. A fuller interpretation is to be given a little later. For the present, it is sufficient to note that ordinarily both the MA and IQ of a pupil should be recorded, if available, for each has its distinctive values—and limitations.

Advantages of the MA concept. The MA has certain outstanding values. Probably the chief of these is that it makes possible a comparison with achievement scores also expressed in age units, as well as with the CA of the pupil, so long as the derived scores are obtained for the same population or from comparable populations. The age basis of comparison is a much more stable unit than the grade location, which is greatly influenced by the promotion policies of the school.

Limitations of the MA. There are also certain serious limitations of the MA, most of which apply particularly to the use of the concept in the high school. It has often been pointed out that the definition of MA does not hold true for CA's beyond 13 or 14. One reason for this is that the norms were based primarily upon pupils in school, who became an increasingly select group, the weaker ones tending to drop out. This was especially true when Terman was standardizing the original (1916) Stanford-Binet scale, against which many later tests were validated. Then, too, in spite of their appearance, the mental age units on the scale are probably of unequal length, the annual increments becoming smaller and smaller as they approach maturity, when the curve flattens out altogether. But no way has been devised so far for equating these units, or for making satisfactory allowance for their variation in length. This is the principal reason why true growth curves of mental development are not obtainable up to the present even when the same individuals have been measured repeatedly over a long period of years. This also complicates the problem of investigating the constancy of the IQ, and of its computation in the later chronological ages. Neither of these limitations, however, is very serious in the elementary school.

There is a more serious limitation which appears to operate on all age

levels: namely, that the mental age units on one test are not fully comparable to those on another test. It is, of course, entirely possible that imperfect standardization is largely responsible. But whatever the explanation, it is clearly necessary in reporting intelligence test scores to indicate both the name of the test and the form used.

It may, of course, be true that under the circumstances the terms MA and IQ are not particularly fortunate when used to describe the scores on existing tests. Be that as it may, the users of these tests should frankly recognize such limitations as exist. It is a curious fact, however, that people are just as loath to recognize the limits of their brain children as are parents to recognize the limits of their flesh and blood children. The difficulty of arriving at a rational interpretation of a low score on an intelligence test may as well be due to myopia on the part of the interpreter as to that condition in the parent whose child received the score. Boynton¹⁴ recommends what he calls a "pragmatic attitude" toward the tests, for the facts are that "in a vast majority of cases they work with a high degree of success."

After all, in spite of certain definite limitations, intelligence tests, when intelligently used, do afford valuable information to classroom teachers and school administrators. So long as that is true, whether they measure intelligence or something else, whether the age score is really *mental* age or only *personal* age, would appear to be primarily a matter of academic interest only.

One other limitation of MA and all other gross units is that by lumping together many elements they obscure significant differences. Two children of the same CA might have an MA of 8 years and yet be quite unlike. One child might be unusually strong in the linguistic elements of the test but lacking in the more concrete, practical, or common sense elements while just the reverse might be true of the other child. This means that the *pattern* of the test responses, as well as the total or average, must be considered. This, of course, does not mean that the total score has no value but rather that by itself it is inadequate, especially for diagnosis and guidance. The practical suggestion, then, is to consider the pattern as revealed by the profile, as well as the total score, be that an age score or what not. As Thurstone says¹⁵ "Each individual should be described in terms of a profile of mental abilities instead of by a single index of intelligence."

The computation of the IQ. As ordinarily written, the formula for the IQ is $100 \left(\frac{MA}{CA} \right)$. That is, the IQ is the quotient obtained by dividing the mental age of the pupil by his chronological age at the time the test was given. In other words, it is the percentage that the mental age

¹⁴ Paul L. Boynton *Intelligence Its Manifestations and Measurement* pages 231-234 New York: D. Appleton-Century Company, 1933.

¹⁵ L. L. Thurstone *A New Concept of Intelligence and a New Method of Measuring Primary Abilities* *Educational Record* 17:133 Supplement No. 10 October 1936.

is of the chronological age. As a matter of fact, however, the CA used as a divisor is never more than the age at which the test maker assumes mental maturity is reached. Upon the basis of the evidence available in 1916, Terman¹⁶ suggested that a divisor of 16 years be used for all pupils whose CA is 16-0 or above. In the Revised Stanford-Binet, Terman and Merrill¹⁷ suggest this rule:

Up to 13-0 the entire C.A. is counted, beyond 16-0, none of it. The C.A. of a subject who is between the ages of 13-0 and 16-0 is counted as 13-0 plus $\frac{3}{4}$ of the additional months he has lived. This means that a true C.A. of 14 is counted as 13.8, a true C.A. of 15 as 14.4, and a true C.A. of 16 as 15.0, which is the highest divisor used in the computation of the I.Q.¹⁸

This suggestion would appear to be in keeping with the fact that mental maturity is reached gradually rather than abruptly. The age at which it is attained probably varies considerably from test to test. Many writers favor using percentiles or standard scores, rather than IQ's, especially beyond the elementary school.¹⁹

The actual work of computing IQ's can be greatly reduced by the use of tables such as those in Terman and Merrill.²⁰ A preceding chapter emphasized the need for a careful checking of the scoring and totaling of all scores and the obtaining of the MA or other equivalents from the tables of norms in the manuals. There is one other step in computing the IQ, even with the use of tables, that must be watched carefully to insure accuracy. That is the determining of the CA of the pupil. In the lower grades this age score should be taken from the date of birth as shown on the school records, which in turn should be based upon a birth certificate. A young child is likely to put down 9 when he is merely "going on 9," for example. In the upper grades it is usually safe to rely on the pupil's answer as given.

¹⁶ Lewis M. Terman, *The Measurement of Intelligence*, pages 140-141. Boston: Houghton Mifflin Company, 1916.

¹⁷ Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence*, page 68. Boston: Houghton Mifflin Company, 1937.

¹⁸ In symbols the three formulas are

$$\begin{aligned} \text{IQ}_{(\text{CA less than } 13-0)} &= 100 \left(\frac{\text{MA}}{\text{CA}} \right) = \frac{100\text{MA}}{\text{CA}} \\ \text{IQ}_{\text{CA} = 13-16} &= 100 \left[\frac{\text{MA}}{13 + \frac{3}{4}(\text{CA} - 13)} \right] = \frac{150\text{MA}}{65 + \text{CA}} \end{aligned}$$

For examinees 16 years old or older the formula becomes

$$\text{IQ}_{(\text{CA} = 16-0 \text{ or more})} = 100 \left(\frac{\text{MA}}{16} \right) = 6.67\text{MA}$$

These formulas apply *only* to the Revised (1937) Stanford Binet Intelligence Scale, Forms L and V. The appropriate denominators for other tests may be quite different.

¹⁹ The Wechsler Intelligence Scale for Children (WISC) has standard-score IQ's for 33 age levels from 3-0 through 15-11. David Wechsler, *Manual for the Wechsler Intelligence Scale for Children*, pages 27-59. New York: The Psychological Corporation, 1943.

²⁰ Lewis M. Terman and Maud A. Merrill, *op. cit.*, pages 417-450.

on the test blank. On most tests he is asked to give his age at his last birthday, and then to give the year, month, and day of his birthday, or else to tell how many months it has been since his last birthday. The trouble usually comes with computing the months. This computation should always be checked, preferably by a simple table prepared by the examiner. Table 31 illustrates such a table, prepared for a test given on May 21, from which the months can be read directly. It is desirable to verify the years

TABLE 31

A TABLE FOR COMPUTING MONTHS SINCE LAST BIRTHDAY
(DATE OF TEST MAY 21)

<i>Birth days Between Dates</i>	<i>Months Since Birthday</i>
January 6 and February 5	4
February 6 and March 5	3
March 6 and April 5	2
April 6 and May 5	1
May 6 and June 5	0 May 21 Test Date
June 6 and July 5	11
July 6 and August 5	10
August 6 and September 5	9
September 6 and October 5	8
October 6 and November 5	7
November 6 and December 5	6
December 6 and January 5	5

for those pupils whose birthdays come in the month the tests are given or in the next month or so, for even high-school pupils will often make an error of one year. When the correct MA's and CA's are determined, the IQ values can be read from a table. If the IQ's are computed by actual division, it is necessary to have the work done twice independently.

Interpretation of the IQ. The IQ is a measure of *brightness*, or of *rate of intellectual development*. Following the lead of Terman, many writers consider IQ's of 90 to 110 as "normal," those below as subnormal and those above as supernormal. According to this scheme IQ's below 70 may indicate "feeble-mindedness." Individuals in this group are often subdivided into three types or levels of feeble-mindedness: idiots, below 25; imbeciles, 25 to 49, and morons, 50 to 69, inclusive. Most clinicians recognize these as rather rough and arbitrary groupings and attempt to apply other criteria, such as social sufficiency or success in school. Feeble-mindedness is a psychological, social, medical, legal concept.

There is a continuous distribution from the idiot to the genius, and the various degrees of brightness shade into each other until they are as indistinguishable as the borderline between colors of the rainbow. It is easy to see that red is different from violet or to see the difference between red

and yellow, but it is hard to tell where orange leaves off and becomes red on the one hand or yellow on the other. The concept of "genius" is worthy of further consideration. Following the lead of Terman it has been common to interpret any IQ of 140 or above as indicating "genius or near genius." Evidence is accumulating which indicates that this limit is much too low. In an illuminating discussion of this problem, Hollingworth comes to the conclusion that a minimum IQ of 170 or 180 is more defensible, and that works of genius are conditioned by high ability when combined with zeal and hard work.²¹ Terman supports the conclusion that "above the IQ level of 140, adult success is largely determined by such factors as social adjustment, emotional stability, and drive to accomplish."²²

Advantages of the IQ. The identification of various degrees of brightness is one of the advantages of the IQ. Moreover, many studies have shown the IQ to remain relatively constant under ordinary conditions from year to year, although radical changes in the home and school environment, which rarely occur, are likely to be reflected in larger changes in IQ when they do occur. Nemzek²³ summarized 97 studies which used the 1916 Stanford Binet test and 27 studies which used group tests. The median correlation coefficient by the test and retest method was .832 for the individual test and .846 for the group tests. The corresponding range of the middle 50 per cent of the coefficients was 760 to 889 and 779 to 880, respectively. The similarity of results for the individual and group tests is remarkable. But these correlations permit considerable variations, which apparently are more likely to occur at the extremes of the distribution than near the center.²⁴ There seems to be a tendency for the lower IQ's to decrease somewhat on later tests, while the evidence for the higher IQ's is somewhat contradictory. After six years Terman found that 73 of his "geniuses," still below a CA of 13, had lost in Stanford Binet IQ, the boys 3 points and the girls 13 points on the average. On the other hand, Cattell at Harvard found that children with IQ's above 120 gained approximately 8 points in three to six years. Most studies have noted a regressive effect however. But, of course, most cases tend to cluster rather closely about

²¹ Leta S. Hollingworth, *Children over 180 IQ: Stanford Binet Origin and Development* (Yonkers-on Hudson, N. Y.: World Book Company, 1942).

²² *Thirty-ninth Yearbook of the National Society for the Study of Education*, Part I, page 81. Bloomington, Illinois: Public School Publishing Company, 1940. Quoted by permission of the Society.

²³ Claude L. Nemzek, 'The Constancy of the IQ,' *Psychological Bulletin* 30, 1-4 February 1933. A later review is Robert L. Thorndike, 'Constancy of the IQ,' *Psychological Bulletin* 37, 167-186, March 1940. For a claim that IQ's are extremely stable from the first grade to the freshman year of college see G. L. Brown, 'On the Constancy of the IQ,' *Journal of Educational Research* 44, 1-11, 1-3 October 1950. A critical rejoinder is Julian C. Stanley, 'A Note Concerning Brown's On the Constancy of the IQ,' *Journal of Educational Research*, 48, 545-547, March 1955.

²⁴ Both statistical evidence and clinical experience seem to agree that the Revised Stanford Binet is particularly reliable at the lower IQ levels. See Quinn McNemar, *The Revision of the Stanford Binet Scale*, page 13. Boston: Houghton Mifflin Company, 1947.

the center of the distribution, where the IQ is "fairly stable." After a summary of the experimental evidence, Cattell arrives at this practical conclusion:²²

The results are reported as evidence of the large changes in the IQ which do occur in ordinary school practice and to emphasize the caution with which the results of a single intelligence test must be interpreted even though it be an individual examination made by an expert.

Since the IQ of the average pupil is likely to be relatively stable if originally computed for CA's between 7 and 13 years his MA at a later age can be estimated with fair assurance from his present CA and his IQ. For example, suppose that a pupil who had an IQ of 95 when in the third grade is now in the fifth grade. His present CA is 10.2 or 10.17 years. His estimated MA in years is 10.17×95 which is $9.66 = 9\frac{2}{3} = 9\frac{8}{10} = 9-8$. Comparisons with achievement test scores expressed in ages can therefore be made wherever such comparisons are thought desirable without the necessity of repeating the intelligence tests at the same time. Although it is desirable to repeat intelligence tests until at least three tests have been given during the pupil's educational career, the tests need not be given at the same time as the achievement tests in order to make comparisons. The IQ from a test given by an expert to pupils in the public school may be regarded as sufficiently constant to make adjustments for a different date fairly safe, at least for a period of two or three years.

Limitations of the IQ. In common with all units in which test scores are expressed, the IQ suffers from two limitations. The zero point is arbitrary rather than real, and the various units are of unequal length or value. The difference between 60 and 70 is not equal to the difference between 90 and 100, or to the difference between 120 and 130. In the same way it is absurd to say that a pupil whose IQ is 120 is twice as bright as one whose IQ is 60, or half again as bright as one whose IQ is 80. But in this regard the IQ is no worse than are all raw test scores and practically all other derived scores. For example, it is obvious that when the thermometer registers 10 degrees below zero it is not twice as cold as when the thermometer registers 5 degrees below.

There is also another serious limitation of the IQ. Many studies have shown that the IQ's on one test are not comparable to those obtained on another test. This was shown clearly in Table 24 on page 110 where the mean IQ's on five intelligence tests ranged from 96.4 to 118.2 for the same 284 high school seniors. The ranges of individual scores on the five tests are not reported, but it is safe to infer that some of them must have been quite large. IQ's from different tests must be equated in order to make them comparable, for even when average IQ's on different tests are close together the extremes are likely to vary widely. One solution which has

²² Psyche Cattell, *Stanford Binet IQ Variations* (School and Society) 4: 610-618 May 1, 1937.

TABLE 32

TABLE FOR EQUATING INTELLIGENCE QUOTIENT VALUES²²(Use only with cases which were sixteen years
of age or over when tested)

Cor- rected IQ Value	IQ on Otis Higher, Form A	IQ on CTMM*			IQ on Terman- McNemar	IQ on SRA PMA	IQ on SRA Non- Verbal	Cor- rected IQ Value
		Total	Non- Language	Language				
90	88	96	87-88	96	87	75	97	90
91	89	97	89-90	97	88	76	98	91
92	90	98	91-92		89	77	99	92
93	91	99	93-94	98	90	78	100	93
94		100	95-96		91	79	101	94
95	92	101	97-98	99	92	80	102	95
96	93	102	99-101		93	81	103	96
97	94	103	102-105	100	94	82-83	104	97
98	95	104	106-108		95	84	105	98
99	96	105	109-110	101-102	96	85-87	106-107	99
100	97	106	111-112	103-104	97	88	108-109	100
101	98	107	113	105	98	89	110-111	101
102	99	108	114-116	106	99	90	112	102
103	100	109	117-118	107	100	91	113	103
104	101	110-111	119		101	92		104
105	102	112	120	108	102	93	114	105
106	103	113	121-122	109	103	94	115	106
107		114	123		104	95	116	107
108	104	115	124	110				108
109	105	116	125-126	111	105	96	117	109
110	106	117	127-128	112-113	106-107	97	118-119	110
111	107	118	129-130	114	108	98	120	111
112	108	119	131-133	115	109	99	121-122	112
113	109	120	134-136	116-117	110	100-101	123-124	113
114	110	121	137-139	118	111	102-103	125-126	114
115	111	122	140-141	119	112	104	127	115
116	112	123	142-143	120	113-114	105-106	128	116
117	113	124	144	121	115	107	129	117
118	114	125	145		116	108	130	118
119	115	126	146	122	117	109	131	119
120	116	127-128	147	123-124	118	110-111	132-133	120
121	117	129	148	125-126	119	112-113	134-135	121
122	118	130	149	127-128	120	114	136-137	122
123		131		129	121	115	138-139	123
124	119	132	150	130	122	116	140	124
125	120	133	151	131	123	117	141-142	125

* These values apply only to the California Test of Mental Maturity, Short Form, 1942 edition

²² This is a slight modification of Table XXIV in an unpublished report by Walter G. Heil and Alice Horn, "A Comparative Study of the Data for Five Different Intelligence Tests Administered to 284 Twelfth Grade Students at South Gate High School—Los Angeles" Los Angeles Curriculum Division, Los Angeles City School Districts, February, 1950 (Mimeographed)

been proposed is to equate all tests in terms of some widely used test.²⁷ Another procedure is to equate several tests locally. This was done by Heil and Horn, as shown in Table 32, for the 284 second semester twelfth-grade students in a Los Angeles, California, high school. Their five intelligence tests were the Otis Self-Administering Test of Mental Ability, Higher Examination, Form A, the California Short-Form Test of Mental Maturity, 1942 edition, the Terman-McNemar Test of Mental Ability, the Science Research Associates Primary Mental Ability tests, and the SRA Non-Verbal form. The bold-face IQ's going from 90 to 125 on each side of Table 32 are corrected values. Note, for instance, in the 117 row that actual IQ's on the various tests range from 107 to 144. At 125 they run from a low of 117 to a high of 151.

Lennon²⁸ has provided tables yielding equivalent scores and equivalent IQ's for the Otis Quick-Scoring Mental Ability Tests Gamma Test, Form A or B, Pintner General Ability Tests Verbal Series, Advanced Test, Form A or B, and Terman-McNemar Test of Mental Ability, Form C or D. An IQ of 100 on the Otis is equivalent to 99 on the Pintner and 102 on the Terman-McNemar. Corresponding to an Otis IQ of 130 is a Pintner IQ of 134 and a Terman-McNemar IQ of 138. At the lowest IQ level shown in Lennon's Table III are the following figures: Otis, 76, Pintner, 64, and Terman-McNemar, 66. Therefore, among these three tests issued by the same publisher discrepancies at the "average" IQ of 100 are slight, but differences may be serious for low or high scorers unless the recommended conversion is made. The Heil-Horn and Lennon studies point up the fact that an individual may have as many IQ's as there are different intelligence tests.

Doubtless the fundamental solution is for all test makers to standardize their tests, whether they aim to measure intelligence or achievement, on a national population so chosen as to conform fully to the mathematical theory of sampling. As long as tests continue to be standardized on samples chosen primarily upon the basis of convenience, even when they involve large numbers and wide geographical areas, there is still no assurance that the samples are truly representative and thus comparable with each other.

Because of its numerous limitations some authorities would abandon the IQ concept altogether. Stoddard,²⁹ for example, characterizes it as a "myth" pure and simple. No one recognizes the limitations of the IQ more clearly than its friends, as this statement from Terman³⁰ indicates: "An obtained

²⁷ W. S. Miller, "Variation of IQ's Obtained from Group Tests," *Journal of Educational Psychology* 24: 468-474, September, 1933.

²⁸ Roger T. Lennon, "A Comparison of Results of Three Intelligence Tests," *Test Service Notebook No. 11*, Yonkers, N. Y.: World Book Company, 1951, 4 pages.

²⁹ George D. Stoddard, *The Meaning of Intelligence*, page 258, New York: The Macmillan Company, 1943.

³⁰ *Thirty-Ninth Yearbook of National Society for the Study of Education*, op. cit. page 466.

I Q is not only subject to chance errors resulting from inadequate sampling of abilities, but also of numerous other errors, including practice effects, negativism or shyness, the personal equation of the examiner, and standardization errors in the test used²¹

All things considered, the authors are disposed to agree with Terman and Merrill²¹ that the sensible thing to do under the circumstances is to "employ the simplest indices available and as rapidly as possible acquaint teacher, school counselors, social workers, and physicians with their significance and their limitation" The MA and the IQ are examples of such "simple indices" However, amateur test users will do well to remember at all times Hildreth's warning that "no one I Q ever indicates exactly any child's tested ability"²² No matter how obtained, the IQ should never be accepted as the final verdict, but rather as a point of departure for further investigation

Other derived scores. To avoid the difficulties in the MA and IQ, other types of derived scores have been proposed Of these, the three most common will be considered briefly It must be apparent at the outset, however, that no norm can be any better than the sample upon which it is based or the measuring instrument employed Errors of sampling and of measurement cannot be avoided by the simple device of shifting the unit in which to express the norms The Cooperative tests have moved in this direction by taking as a point of reference the "50 point," the score "intended to represent the score which the average white child in the United States would make at the end of the particular course if he had attended a typical school and had had the usual instruction in the subject in question"²³

The Personal Constant (PC) has been suggested by H Heims as a substitute for the IQ Kuhlmann was so convinced of the merits of this method that he included a table of Heims Mental Growth Units, which he recommended in place of the IQ for use with the Kuhlmann-Anderson tests On the other hand, Cattell²⁴ found the PC more constant than the IQ for pupils of low intelligence but not for those of high intelligence The PC has not received wide acceptance

A second substitute, proposed by many writers, is the *percentile rank* A percentile rank is a description of a pupil's position in a typical age or grade group in terms of the percentage of pupils who fall below that score A percentile rank of 50 would, of course, be exactly at the median In like

²¹ Lewis M Terman and Maud A Merrill *op cit.* page 29

²² Gertrude Hildreth "Stanford Binet Retests of Gifted Children," *Journal of Educational Research* 37 301 December 1943

²³ John C Flanagan *The Cooperative Achievement Tests A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores* page 19 New York Cooperative Test Service 1939

²⁴ Psyche Cattell "The Heims Personal Constant as a Substitute for the IQ" *Journal of Educational Psychology* 24 221 228 March 1933

manner a percentile rank of 10 would show that in a typical group only 10 per cent make a poorer score than that, while a percentile rank of 90 would mean that only 10 per cent make better scores than that, since 90 per cent fall below. This is a very simple and useful system that is widely used for achievement tests also, but it has two limitations. One is that a percentile rank of a given magnitude in one group is not directly comparable with the same percentile rank in another group. A 10th percentile pupil in the freshman class is manifestly not the same as a 10th percentile pupil in the senior class, for example. A second limitation is that the percentile rank units are of unequal length. For example, in a typical group an IQ of 62 has a 1st-percentile rank and an IQ of 68, which is an increase of 6 points, has a 3rd percentile rank, but another IQ change of 6 points, from 81 to 91, raises the percentile rank 11 points. In other words, the distances between percentiles near the center of the group are much less than those at the extremes.

A third procedure is the method of *standard scores*, sometimes called sigma scores or *z-scores*, used by Stutsman in her Merrill Palmer performance scale and by Wechsler in his scales. These units are expressed in terms of the mean and standard deviation of the typical age or grade group or, for that matter, of any group. An illustration will help to make the system clear. Suppose that a pupil makes 40 points on one test and 80 points on another test. It is clearly unsafe to say that he did better on one test than on the other. This is evident if we find that the mean score on the first test is 30 points and that the mean score on the second is 90 points. In other words, the pupil is 10 points above average on one test and 10 points below average on the other test. To reduce these scores to a common denominator requires one additional step, namely, to take into consideration the variability of the scores as well as their central tendency. Suppose then, that the standard deviation of the first is 10 points and that of the second test is 20 points. It is now clear that our pupil is 1.0 standard deviation distance above the mean on one test and .5 standard deviation distance below the mean on the other. These two figures are standard or sigma scores and are written $+1.0\sigma$ and $-.5\sigma$. To avoid negative numbers, the suggestion is sometimes made that the mean score be called 50 arbitrarily and each standard deviation distance above and below be equivalent to 10 points. In our illustration opposite, the pupil's scores would be $50 + (1 \times 10) = 60$, and $50 - (.5 \times 10) = 50 - 5 = 45$.

The system has much to commend it statistically. In fact its principal limitation is that it appears to be rather cumbersome to handle. That impression is, however, probably due more to its unfamiliarity than to any thing else. Some writers³⁵ point out that not only are MA's and IQ's defined in the usual fashion indeterminate for the upper half of the adult popula-

³⁵ L. L. Thurstone and Thelma Gwinn Thurstone, *Psychological Examinations* 1940
Norms. American Council on Education Studies 5 2-3 1941

tion, but they also argue that standard scores or percentile scores yield much more information even for young children. Figure 35 shows the relation between standard scores, percentile scores, and Revised Stanford-Binet IQ's. It will be noted that the IQ's on the Revised Stanford-Binet may be considered roughly as standard scores whose mean is 100 and whose sigma is 16.

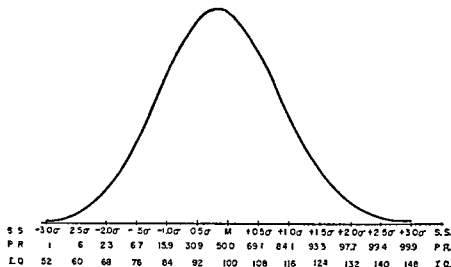


Figure 35 The Relation Between Standard Scores, Percentile Ranks, and Revised Stanford-Binet IQ's (Based on Terman and Merrill, *Measuring Intelligence*, page 42)

Regardless of the type of norms used, the teacher must never lose sight of the fact that all measurement is subject to error and that scores can rarely be taken at face value. Some persons are so impressed by the "ubiquitous probable error," to use Kelley's phrase, that they think numerical scores of every kind "convey an unwarranted impression of exactitude," and would report the results of intelligence tests in general terms, such as "dull," "normal," or "bright."³⁶ In the writers' judgment, a better practice is to continue to employ the numerical scores but to be keenly aware of their limitations.

D. The Use of Norms in Interpreting Scores on Achievement Tests

Educational age versus educational quotient. In interpreting scores on achievement tests the terms *educational age* and *educational quotient* are sometimes used in just the same way that *mental age* and *intelligence quotient* are used in interpreting scores on intelligence tests. In other words, educational age, or EA, is a measure of educational maturity, or level or stage of educational growth. In like manner the *educational quotient*, or EQ, is a measure of rate of educational growth or development. The EQ is

³⁶ H. F. Garrett, 'The Standardization of the Terman-Merrill Revision of the Stanford-Binet Scale,' *Psychological Bulletin* 40:196, March, 1943. Also see *Psychological Bulletin* 43:72-76, January 1946.

obtained by dividing the EA by the CA. For example, a 10-year old boy has made a score of 60 points on a certain achievement test, which is the average score for a 12-year old pupil. The boy is then said to have an EA of 12-0, which, divided by 10-0 gives him an EQ of 120. In like manner another 10-year-old boy in the same class might make a score of 35 points, which is the average score for a pupil of 8 years and 6 months. His EA is 8-6, and his EQ is 85. It should be noted that the terms EA and EQ refer to scores made on general achievement tests or on test batteries involving several subjects. If a test in only one subject is used the terms *subject age* and *subject quotient* are employed. For example, a reading test would yield reading ages and reading quotients, while an arithmetic test would yield arithmetic ages and arithmetic quotients and so on for the other subjects.

Uses of EA. The value of EA and of the various subject ages is that they make possible a meaningful interpretation of scores in terms of a relatively stable unit, chronological age. They also facilitate important comparisons with norms, on both intelligence and other achievement tests, whenever they have been standardized on comparable groups, as well as with the individual's own MA and CA.

Limitations of EA. EA and all subject ages have many of the limitations already pointed out in the case of the MA. Probably the most serious is that they reflect the promotion policies and holding power of the schools in which the tests are given. It is a matter of common observation that the performance of a 10-year-old pupil who is retarded in the grade is not the same as that of a 10-year-old pupil who has made normal progress, and much less than that of the accelerated pupil of the same age. Crawford²⁷ has made an extensive study of the influence of such factors upon norms based on unselected groups, and comes to this significant conclusion: "The factors of chronological age, mental age, and rate of progress affect test norms to a degree that makes the use of norms based on groups in which these are not controlled of doubtful value." His recommendation is that both CA and MA should be used in establishing norms. One solution to the problem is to use only pupils whose CA's are normal for their respective grades in computing norms.

One other limitation of existing age norms is that age units on one test are not comparable to those on other tests that are presumably measuring the same thing. Test publishers owe a service to the public and should prepare tables for equating age norms on various achievement tests in much the same way as has been done by the Cooperative Achievement Tests²⁸ and the various parts and forms of the Metropolitan and Stanford Achievement Tests.²⁹ Another less serious limitation is that the age units on

²⁷ John R. Crawford, "Age and Progress Factors in Test Norms," *University of Iowa Studies in Education*, 9, 1, 39, June 15, 1934.

²⁸ Published by the Cooperative Test Division of Educational Testing Service.

²⁹ Published by World Book Company.

any one scale or test are not necessarily equivalent throughout its length

An important limitation of the EA and all other gross units is that they lump together many and diverse elements in such a way as often to obscure significant differences. Two pupils of the same CA or MA might have an EA of 10-0, for example. This does not guarantee that they are by any means identical in achievement. One pupil might be greatly accelerated in reading, language, and literature, but retarded in arithmetic, spelling and science, whereas the exact opposite might be true of the other pupil. EA, which is a composite, or average, has taken no account of the *pattern*, which may afford the key to an adequate interpretation. This fact, of course, does not mean that age scores and other averages have no value, but rather indicates that they are inadequate by themselves. The practical implication is clear. The total score, whether age or what not, is important, but must be considered always in relation to the *pattern* of the responses, usually best represented as a profile. Figure 36 showing two profiles drawn

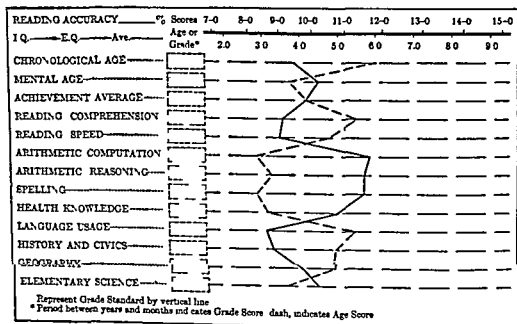


Figure 36 The Profiles of Two Pupils Who Made the Same Total Score on a General Achievement Test (From the Modern School Achievement Tests published by Bureau of Publications Teachers College Columbia University)

upon the same chart for pupils making the same total score, should make this point clear

Use and limitations of EQ The method of computing the FQ and the various subject quotients has already been described. As measures of rate of educational progress these quotients may be useful in the interpretation of scores on achievement tests. However, no such elaborate scheme for the interpretation of these quotients as was described for the interpretation of quotients on intelligence tests has been worked out. There is

nothing which corresponds to such terms as "feble-mindedness" or "genius"

EQ's are subject to some, but not to all, of the limitations pointed out for IQ's. Since education and growth continues at least throughout the formal school period, there is no problem of selecting a maximum divisor such as was described in the case of IQ. It is true that the units are of unequal length and that the quotient technique is not appropriate for use in high school, where age norms on achievement tests are ordinarily not available. Undoubtedly the most serious limitation of quotients, as well as of other norms, is that the units on one test are not directly comparable with those on another test that purports to be measuring the same thing. Tables for equating quotients on achievement tests similar to those for intelligence tests have not been published. Better still would be the exercise of greater care in the original standardization. The test record should always indicate the name of the test and the form used, for achievement tests as well as for intelligence tests.

Use and limitations of grade norms. It is a very common practice to interpret achievement tests in terms of *grade norms*. Grade norms on standard tests are usually the average scores made on the test by pupils in each grade when the test has been given to pupils in widely scattered areas. In the earlier tests these grade norms were usually for the end of the grade only, although sometimes for the middle of the grade also. This made comparison with norms somewhat difficult, unless the tests were given at the same time in the year. Figure 31, on page 267, illustrates a simple graphical method of making comparisons with norms for a different time in the year. Of course, such a comparison assumes uniform progress throughout the grade, which may be only approximately true in some instances. A slight variation of such norms for high school use is to base the norms upon the length of time the subject has been studied rather than upon the grade or year in which it happens to be offered. Since many high-school subjects do not continue over the entire high-school period and have no definite grade location, norms based on the number of semesters the subject has been studied are very useful. The problem of interpretation is just the same as for regular grade norms.

In recent years many tests below the senior high school level have norms available for every month in the school year. For example, 6.0 means the norm for the beginning of the sixth grade, while 6.5 is the norm for the middle of the grade. In like manner 4.2 means the norm for the fourth grade two months after school starts, and 4.10 means the norm for the end of the fourth grade. Such norms are often called *G scores*, and sometimes *B-scores*. They have the distinct advantage of being readily understood. They also have certain dangers and limitations. For one thing they tend to imply a degree of mathematical exactness which the accuracy of existing tests hardly warrants. Certainly it is unsafe to take them literally at their

face value. A still more serious limitation is the lack of comparability of scores on different tests. Adams found,⁴⁰ for example, that eight arithmetic tests rated the mean performance of 152 pupils all the way from the fifth grade to the eleventh grade, depending upon the test used. It is unnecessary to comment upon the absurdity of fractional grade norms in a situation like that. The solution is, however, not so much in the abandonment of grade norms as in their further refinement. There is another danger in interpreting grade norms, no matter how accurately determined. This danger arises partly from the fact that a school with an overstrict promotion policy will tend to show up favorably on grade norms simply because of the presence of a great many pupils in the several grades who really belong in higher grades. It is always well, therefore, in any apparently superior school, to make a comparison on the basis of age, to see whether the superiority is real or only illusory. Of course there would be little difference between schools which promote strictly on the basis of CA and those in which the percentages of acceleration and retardation are balanced, a condition which rarely exists.

Ruch and Segel, after noting some evidence that "recent tests may possibly have much more dependable norms than those standardized a decade or so earlier," nevertheless make the suggestion that⁴¹

many factors peculiar to the individual school system must be considered in the interpretation of tests such as the legal age of school entrance, the actual average age of school entrance, rates of acceleration and retardation, rates of elimination from school, percents of failures of pupils, genuine differences in instructional efficiency, and variations in average mental and educational capacity from school to school.

Harris⁴² has called attention to a rather common error made in the interpretation of grade norms at the primary level. It arises from the failure to take into account the fact that zero performance on an achievement test is 1.0. A first grade class whose grade score at the end of the year is 2.0 has made only normal progress for the year.

Other norms for achievement tests. Several other types of norms are used, some of which require brief mention. Of these, doubtless the most important are *percentile norms*. As in the case of intelligence tests already discussed, such norms interpret a pupil's score by describing his position in the group in terms of the per cent of pupils who fall below the score made. Generally, all percentile ranks from 0 through 99, but sometimes only certain points such as the 25th, 50th, and 75th, are given. These percentile

⁴⁰ Summarized by Giles M. Ruch from an unpublished master's thesis by Eunice Adams, *The Comparative Reliability of Eight Arithmetic Tests*, University of California, 1929, in *Review of Educational Research* 3: 39, February, 1933.

⁴¹ Giles M. Ruch and David Segel, *Minimum Essentials of the Individual Inventory in Guidance*, page 82, Washington: United States Office of Education, 1939.

⁴² Albert J. Harris, "Note on a Source of Error in Interpreting Grade Norms," *Journal of Educational Research* 39: 151-153, October, 1940.

ranks are very easy to interpret, but they have two limitations, neither of which is usually very serious for most purposes. One is that the scale values are unequal in length, and the other is that percentile values in one grade or age group are not directly comparable with those in another.

Standard scores are also used. They are interpreted in the same manner as are similar scores on intelligence tests. These have already been discussed. McCall has proposed a modification called a *T-score* based upon a standard group composed of 12-year-old pupils. All age and grade groups are described by locating their *T-score* position in this 12-year old group. The mean is given a value of 50, and each standard deviation distance above and below is divided into tenths, each counting one point. For example, a 15-year-old pupil makes a reading score on the Thorndike-McCall Reading Test, which, according to the table of norms, is a *T-score* of 60. In other words, this pupil is located 1 σ distance above the mean of typical 12-year-old pupils who have taken the test. The *T-score* technique is sometimes used with other age groups. The principal limitations of the *T-score* are that it is not well adapted to high-school tests and that it is rather cumbersome even for grade-school tests.

Value of local norms. Practically all norms published on tests are so-called *national norms*. When such norms are carefully derived, they are of great value in interpreting the scores. It is easy to overemphasize their value for the ordinary school and school system, however. They must never be taken as standards. There are such wide variations in the length of school terms, in the equipment of schools, in the training and experience of teachers, and in other important respects among the several states and among the school units of any one state as to make any single series of norms for the whole nation inadequate. National norms must be supplemented by norms for the state, county, and city school systems, and even for the individual school. What is really important in most cases is the comparison of grades, classes, and schools which operate under approximately the same conditions. Lindquist⁴³ has pointed out several distinct advantages of the regional testing programs used in Iowa for a number of years.

For purposes of classification, what is needed is a set of norms for the school itself. To derive satisfactory local norms, all that is required is to combine all pupils in the same grade and then to compute standard scores or percentiles. If age norms are desired, the pupils will be distributed according to CA or MA, and the medians computed. In larger schools and school systems norms should be derived for slow and rapid learners as well as for average or normal learners on each grade level.

It would appear, then, that the more specific the norm the more useful

⁴³ E. F. Lindquist "Nationally Coordinated Regional Testing Programs in *New Directions for Measurement and Guidance* pages 87-103. Washington: American Council on Education, 1944.

it becomes Educators are coming to recognize that each individual has his own unique pattern of growth This position is clearly stated as follows ⁴⁴

The time has come when we should cease to be primarily interested in comparing one child with another, one class with another or any class with a norm We should be primarily interested in comparing each child with himself with his past record and with his potentialities To center attention elsewhere is to miss the point—to miss the service which tests can render

Figure 37 is the test score profile of a college junior, Richard Roe, based upon local norms at "Siwash" College He was administered an intelligence test, an English battery consisting of grammar, organization, and reading tests, the latter having vocabulary, reading speed, and reading level subtests, and a five test achievement battery history, literature, science, art, and mathematics The norms are shown both as percentile ranks and as "stanines" (standard scores on the basis of 9 points—that is, running from 1 through 9) *TE* means "total English score," and *TA* means "total achievement score" The intelligence, *TE* and *TA* points are connected with a solid line, while the battery tests and subtests are joined by dotted lines

In consultation with his adviser, Richard can see at a glance that he is above the Siwash junior class average in general, but that his history, literature, and art scores fall considerably below the other eight points Richard and his adviser can use this information to good advantage in planning his last two years of college courses

E. Methods of Comparing Intelligence and Achievement

One of the most important questions to raise about any pupil is How well is he getting along in comparison with his capacity? Whenever intelligence tests and achievement tests for the pupil have been expressed in comparable terms a rough answer to this question is possible But the problem is far more difficult than would appear on the surface According to Kelley,⁴⁵ about 90 per cent of whatever is measured by a so-called general intelligence test is the same as that measured by an all around achievement test battery In like manner he has computed the "community of function" between intelligence tests and arithmetic tests as about 88 per cent and that between intelligence and reading tests as about 92 per cent Since a "scant one-tenth" of the tests are utilized in the measurement of difference between intelligence and achievement, he points out the serious hazards of such comparisons It would certainly appear that unless the

⁴⁴ Douglas E. Scates "The Improvement of Classroom Testing" *Review of Educational Research* 8:532 January 1939

⁴⁵ Truman Lee Kelley *Interpretation of Educational Measurements*, page 208 Yonkers World Book Company 1927

The accomplishment quotient. In 1920 Franzen⁴⁷ suggested the *accomplishment quotient*, abbreviated AQ. This is the ratio between EA and MA, or between EQ and IQ. The simplest formula is $AQ = 100(EA - MA)$. A quotient of 100 is considered the goal. For example, if a pupil whose EA is 9-2 has an MA of 10-0, his AQ is $9\ 17 - 10\ 00 = 91\ 7 = 92$. In like manner, a second pupil might have the same EA, 9-2, but an MA of only 8-3. His AQ would be $9\ 17 - 8\ 25 = 111$. The interpretation of the first case, 92, is that the pupil is not living fully up to his capacity, which seems reasonable enough, human nature being what it is. But the interpretation of the second case, 111, is rather absurd, since it appears to imply that this pupil has exceeded what he is capable of doing by 11 per cent! A more probable explanation is that the quotient is due to inaccuracies in the tests and that in this case the errors in the achievement score were in the direction of making it too high, whereas the errors in the intelligence score were in the direction of making it too low. The resulting quotient has added these errors. If the errors, whether due to chance or otherwise, had been in the same direction, they would have tended to offset each other. One reason the use of IQ and EQ involves less risk is that CA, the divisor, is almost wholly free from errors of measurement if obtained from a birth certificate. *Studies by Cureton,⁴⁸ Haggerty,⁴⁹ Tsao,⁵⁰ and others have brought the AQ and similar measures into general disrepute.*

Combining intelligence and achievement scores. Several proposals have been made for combining scores on intelligence and achievement tests, usually for purposes of pupil classification. One of the simplest of these proposals is to average the pupil's rank on the two tests. More refined methods involve the use of some common denominator, such as the standard score. One publication⁵¹ suggests the use of *promotion age* and *promotion quotient* as a basis of classification for instructional purposes. Promotion age (PrA) is the average EA and MA. In this average the two ages may be weighted equally or unequally, whichever seems best for the data in hand. Then the promotion quotient (PrQ) is the $PrA - CA$. On the face of it, such practice appears to be averaging things as unlike as cattle and horses. But, if Kelley's point regarding the great community of function between intelligence and achievement tests is well taken, the practice would appear to be justified on theoretical grounds. And if it provides a better

⁴⁷ Raymond Franzen. 'The Accomplishment Quotient.' *Teachers College Record* 21: 432-440. November 1920.

⁴⁸ Edward E. Cureton. 'The Accomplishment Quotient Technic.' *Journal of Experimental Education* 3: 315-326. March 1937.

⁴⁹ Lida Harmar Haggerty. 'An Evaluation of the Accomplishment Quotient: A Four Year Study at the Junior High School Level.' *Journal of Experimental Education* 10: 78-81. September 1941.

⁵⁰ Fei Tsao. 'Is the AQ or F Score the Last Word in Determining Individual Effort?' *Journal of Educational Psychology* 34: 513-526. December 1943.

⁵¹ *Supervisors Manual for the Metropolitan Achievement Tests*, pages 38-39. Yonkers World Book Company, 1933.

basis for grouping pupils, as appears often to be the case, it has justified itself in practice

F. The Use of Norms in Interpreting Scores on Personality Tests

As a rule, personality test manuals do not contain elaborate systems of norms. The problem is inherently more difficult than that presented by either intelligence or achievement tests, for which we have seen that the norms are far from ideal. Indeed, the very essence of personality is its uniqueness. It is here that the good judgment and common sense of the teacher are highly important.

Terman⁴² strongly questions the possibility, or desirability, of establishing norms for evaluating or adjusting personalities. He says

The psychologist stands aghast at the self assurance with which the professional school counselors in America diagnose the personality faults of little children and at the boldness with which they undertake the delicate task of adjustment. The student of genius who is familiar with the motivating influences that have their origin in quirks of childhood personality shudders to think what the result would have been if school counselors had had a chance to 'adjust' the personalities of the budding geniuses of history. One can imagine them 'reed from all their peculiarities and complexes' adjusted to the world as it was and becoming indistinguishable from the common herd.

On the same point Poffenberger⁴³ quotes with approval this statement from Burbank,⁴⁴ growing out of a lifelong study of plant life

One of the greatest fallacies of near science and of amateurs in Nature's school is the belief that only from the normal can we get our best development and results. As a matter of fact Nature shows us again and again that it is from abnormalities that some of our most valuable and beautiful plants arise. From that weak or abnormal plant—that genius plant—may come the very characteristics that we are looking for, and our only problem is to nurse it physically and keep it strong to pass on its overload of spiritual or esthetic experiences to its children.

Probably the professional educator could hardly do better than to accept wholeheartedly the motto of the founder of the eugenics movement in England, which was 'Treasure your exceptions.' Those who deviate most widely from the average deserve special consideration. It is from this group that geniuses are recruited as well as social misfits of all types. It is socially undesirable, as well as psychologically impossible, to make everybody alike.

A distinguished psychiatrist gives this wholesome comment⁴⁵

⁴² Lewis M. Terman, *The Measurement of Personality*, Science 80: 605-608, December 28, 1934.

⁴³ Albert T. Poffenberger, *Psychology and Life*, Psychological Review 43: 30, January 1936.

⁴⁴ Luther Burbank and Wilbur Hall, *The Harvest of the Years*, page 273, Boston: Houghton Mifflin Company, 1927.

⁴⁵ Karl A. Menninger, *The Human Mind* (Third Edition), page xiv, New York: Alfred A. Knopf, Inc., 1945.

The adjuration to be "normal" seems shockingly repellent to me, I see neither hope nor comfort in sinking to that low level. I think it is ignorance that makes people think of abnormality only with horror and allows them to remain undismayed at the proximity of "normal" to average and mediocre. For surely anyone who achieves anything is, *a priori*, abnormal, this includes, not only the geniuses, but the presidents, the leaders, and the great entertainers. I presume most of the people in *Who's Who in America* would resent being called normal.

As a summarizing statement concerning norms Flanagan's opening sentence is highly pertinent.³⁶

Test scores are meaningful and valuable to the extent that they can be interpreted in terms of capacities, abilities, and accomplishments of educational significance.

SELECTED REFERENCES FOR FURTHER READING

- Chauncey, Henry, and Frederiksen, Norman, "The Functions of Measurement in Educational Placement," Chapter 4 in E. F. Lindquist (Editor), *Educational Measurement* Washington, D. C. American Council on Education, 1951.
- Cook, Walter W., "The Functions of Measurement in the Facilitation of Learning," Chapter 1 in E. F. Lindquist (Editor), *Educational Measurement* Washington, D. C. American Council on Education, 1951.
- Flanagan, John C., *The Cooperative Achievement Tests, A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores* New York: Cooperative Test Service, 1939. 41 pages.
- Flanagan, John C. (Chairman), "Establishing the Type of Norms Most Useful and Important for the Interpretation of Achievement Test Scores," pages 65-113 in *Proceedings of the 1948 Invitational Conference on Testing Problems* Princeton, New Jersey: Educational Testing Service, 1949. Papers by Lee J. Cronbach, Walter N. Duros, Eric F. Gardner, E. F. Lindquist, Robert L. Thorndike and Arthur E. Traxler.
- Flanagan, John C., "Units, Scores, and Norms," Chapter 17 in E. F. Lindquist (Editor), *Educational Measurement* Washington, D. C. American Council on Education, 1951.
- Rulon, Phillip J., "Problems of Regression," *Harvard Educational Review* 11: 213-223, March, 1941.
- Rulon, Phillip J., "On the Concepts of Growth and Ability," *Harvard Educational Review*, 17: 1-9, Winter 1947.
- Seashore, Harold G., and Ricks, James H., Jr., "Norms Must Be Relevant," *Test Service Bulletin No. 39*, pages 1-4. Psychological Corporation, 522 Fifth Avenue, New York 18, New York, May, 1950.
- Stanley, Julian C., "Why Wechsler Bellevue Full-Scale IQ's Are More Variable Than Averages of Verbal and Performance IQ's," *Journal of Consulting Psychology*, 17: 419-420, December, 1953.
- Tiedeman, David V., "Has He Grown?" *Test Service Notebook No. 12* World Book Company, Yonkers-on-Hudson, New York, 1952. 4 pages.

³⁶ John C. Flanagan, "Units, Scores, and Norms," page 695 in E. F. Lindquist (Editor), *Educational Measurement* Washington, D. C. American Council on Education 1951.

PART IV

Measurement in Instruction

II

Motivation and Practice as Related to Testing

A. The Problem of Motivation

Importance of motivation. It is generally recognized in ordinary experience that motivation occupies an important place in human affairs. Such familiar proverbs as "You can lead a horse to water but you can't make him drink" and "It is hard to teach an old dog new tricks" assign to motivation a key position. The horse does not drink for the simple reason that he does not *want* to drink, and the old dog's poor performance is due not so much to lack of ability as to the fact that he has become too well satisfied with the tricks he already knows. In a like manner, every experienced teacher has seen pupils of mediocre capacity succeed because of interest and enthusiasm, while others of more promise have failed utterly because of lack of it. With these observations growing out of ordinary experience the views of psychologists and other keen students of education are in accord.

Meaning of motivation. The term *motivation* is very inclusive. Literally it means *causing movement*. A convenient grouping of motives into two major classes is suggested—internal or organic, and external or environmental. In recent years the term *drive*, or *urge*, has been used for the former, and *goal*, or *incentive*, for the latter. But in the final analysis, motivation, though in some instances externally initiated, always functions internally. Hunger, thirst, and sex, as well as interests, attitudes, wants, desires, and temporary mental sets, are examples of drives. Incentives may be negative as are pain or punishment, or positive, as are rewards in a multitude of forms. A further distinction is often made between motives which are natural or intrinsic, such as a child's interest in play or the

movies, and those which are artificial or extrinsic, such as prizes, marks, grades, credits, and honor rolls

Relation of measurement to motivation. Measurement is related in at least two ways to motivation. In the first place, there is the problem of the measurement of motivation itself. It is often important to know the differences among individuals in the strength of various motives, the comparative strength of the same motive under varying conditions, and the strength of a given motive in comparison with other motives in the same individual. As the development of wholesome attitudes and interests is an objective of modern education, it is just as necessary to know how to measure it as to know how to measure any other objective. While much valuable work has been done in the measurement of animal drives, up to the present no convenient technique has been devised for measuring human motives in any precise manner. The measurement of motivation in education is, then, a problem for the future.

In the second place, there is the problem of the relation of the measurement of educational capacity and achievement to the motivation of learning and teaching. Since teaching and learning are two aspects of the same process, it is reasonable to expect that measurement will be intimately related to both. Some of the more important relationships will be considered in the next two sections.

B. The Relation of Measurement to Motivation in Teaching

Purpose of the teacher and measurement. An obvious relationship of measurement and motivation in teaching arises from the fact that the purpose of the teacher determines the type of measurement used. Whether, for example, the teacher gives many tests or few, long tests or short, informal tests or standardized, survey tests or diagnostic, depends upon his purpose. Since not all tests serve the same purpose equally well, as has been pointed out, the choice of the measuring instrument becomes a matter of primary importance. This problem has already been considered at some length in Chapter 4. Certain points to be raised later in this chapter will also have a bearing upon it.

Teaching emphasis and measurement. The proper teaching emphasis is determined by the results of measurement. Measurement directly demonstrates the quality of the pupil's learning, but it also indirectly reflects the quality of the teacher's teaching. In the light of measured results conscientious teachers attempt as far as possible to correct weaknesses in past teaching and to prevent their recurrence in future teaching. Messenger¹ studied the "influence of the Iowa Academic Testing Program in relation to the teaching of English mechanics in an Iowa high school" and found that the effect had been to "motivate teachers to greater effort" with the allotment of more time to the subject and the use of more drill.

¹ Unpublished master's thesis, University of Iowa, 1931.

material. One of the chief values of measurement may well be its motivating effect upon the teacher.

Taba early realized this relationship and spoke regretfully of the "formidable and serious handicap" of the progressive schools due to the "lack of forms of testing that are in harmony with their aims and adequate to their purposes." Then she added these significant words:²

After all one teaches only what one in some way or another is able to evaluate as an outcome of that teaching. If we are unable to evaluate the growth of integrations and meanings and ways of behavior, we are unable even to form an adequate notion of them still less to guide the process of learning in these terms.

Attention should be called to the fact that Taba's recognition of the limitations of existing measurement does not blind her to the important motivating influence of measurement whether good or bad, and to the urgent necessity for continuous improvement of measuring instruments.

There is, of course, some danger that the content of the examination may exert too great an influence over the teaching emphasis and curriculum content. It has often been alleged that something like this has happened in the case of the New York Regents and the College Entrance Board examinations, where an important effect of such examinations has been to turn secondary schools into "cramming" schools pointing toward the probable content of such examinations as revealed by past examinations by the same agencies. Insofar as this is true, it represents unfortunate and unwarranted control over teaching procedures, which not only defeats the purpose of the examination but places an obstacle in the way of education itself. Such practice fails to take into account the sampling nature of all tests. Any attempt to drill pupils in advance specifically upon the items of the test tends to narrow teaching to the scope of the testing, so that the two tend to become synonymous rather than one a random sampling of the other. This should be clearly understood, and every reasonable precaution should be taken to insure that state-wide testing programs and other forms of academic competition do not promote mere monotonous drill exercises of an especially narrow and vicious type.

A few studies have been made of the effect of exempting pupils from final examinations. An early study by Anderson³ indicated that exempting high school pupils who reached a minimum standard from final examinations had "played havoc with teachers' grades," while the actual performance of the pupils had shown "no appreciable increase." Apparently the motivation had been in the wrong place. That such a disastrous result need not occur is indicated by a subsequent study of the same problem reported

² Hilda Taba, *The Dynamics of Instruction*, p. 18, New York: Harcourt Brace & Company, Inc., 1932.

³ C. J. Anderson, *Is the Exemption System Worth While?* School and Society, 357-360, March 4, 1916.

by White,⁴ who found that the general distribution of marks in his school had changed very little under the exemption system. Even here, however, was found a "decided dip in the distributions immediately below the exemption point of 85 per cent and a corresponding rise just above it." Schools employing an exemption system should remain constantly on guard lest its effect be more to stimulate the teachers to *give* high marks than the pupils to *earn* them.

C. The Relation of Measurement to Motivation in Learning

Close as is the relationship of measurement to the motivation of teaching, the relationship is even closer to the motivation of learning. Elsewhere Ross stated the problem as follows:⁵

Behind the *act* of learning is the *capacity* to learn, and back of the *capacity* is the *motive* to learn—the desire, urge, impulse, drive, or something, that makes the creature *want* to learn, that pushes him out to meet his environment. One of the reasons why most of the correlations of mental capacity with actual achievement, in school and out, have been disappointingly low, has been that students of real ability have not felt a proper urge to work, while those of mediocre talent have frequently possessed the urge to achieve.

If we are ever to be successful in our efforts to predict achievement, therefore, we must not be content with merely analyzing the learning process, understanding the mechanism of learning, its structure and laws of operation, nor with merely exploring the height and range of human possibilities, but we must also find out about the dynamic aspects of human nature. We must discover not only *how* the mind works, but *why* it works when it does and the way it does.

The foregoing statement is an introduction to a report of an experimental attack on one phase of motivation. The study will be briefly presented here as an illustration of one type of psychological experiment concerned with this important problem. Afterwards some critical comments upon this and other similar experiments will be given.

An experiment in motivation. The problem was to determine the influence of a knowledge of results upon the achievement of 59 college students in a simple act of motor skill, making tally marks (N). The procedure was as follows. Upon the basis of an initial practice period three equivalent groups were formed. One of these, the control, had *no knowledge* of its progress, throughout the first ten practice periods of one minute each. During this time one experimental group had *full knowledge* of results, and the other had *partial knowledge*. At the beginning of each practice period after the first, each pupil in the group with full knowledge was shown his paper of the preceding day, with scores and corrections indicated. A distribution of scores for this group was placed on the board, and each student

⁴ Clyde W. White, 'The Effects of Exemptions from Semester Examinations on the Distribution of School Marks,' *School Review* 39: 293-299, April, 1931.

⁵ Clay Campbell Ross, 'An Experiment in Motivation,' *Journal of Educational Psychology* 18: 337-346, May, 1927. Page 337.

was urged to watch his daily progress, both relative and absolute. In the experimental group with partial knowledge of results, each student was told whether he was above or below the average of the group, but that was all. At the end of ten practice periods conditions were reversed; the group which had had no knowledge of results was then given full knowledge for two additional periods, and the other two groups were given no knowledge for these two periods.

The results are shown in Figure 30 on page 266. On the whole, they seemed to justify the conclusion that "the addition of a single other motivating factor, knowledge of results, is sufficient to give the pupils with such knowledge a distinct superiority over the others, and the degree of superiority is roughly proportional to the amount of information possessed."

Limitations of experiments on motivation. The experiment just summarized illustrates several of the weaknesses of the experimental work so far reported on motivation and, for that matter, on other phases of learning as well. There may be conveniently grouped under three headings: factors studied, subjects used, and conclusions drawn.

In the first place, the factors so far studied leave much to be desired. Most of the studies involved are concerned with highly artificial and often trivial tasks. Take the above experiment as an example. It is highly improbable that students will show much enthusiasm over making, as rapidly as possible, groups of four vertical lines crossed horizontally with a fifth. Tallying, to be significant to most persons, would have to be employed as a record of some athletic contest or other situation in which it is a means to an end and not an end in itself. A survey of the literature reveals that a large percentage of motivation studies have been concerned with making legible a's, canceling numbers or letters, assigning a number to a dictated word, learning trivial facts or actual misinformation, running mazes, and the like. What we need to know is how people behave under actual school conditions. Even when arithmetic and other school materials are employed, the experiments are rarely carried on long enough for the novelty of the task to wear off and for the experimental factor to operate under reasonably normal conditions. The total learning time is frequently less than one hour, and often it is only ten to fifteen minutes, as in the above laboratory experiment. To be most helpful in guiding teachers in the day-by-day conduct of their classes these experimental factors must be continued for at least several weeks.

In the second place, the choice of subjects for the experiment has usually been rather unfortunate. Many of the laboratory studies of the effects of rewards and punishments have been limited entirely to animals. No matter how thorough a believer one might be in evolution, one must certainly see that the behavior of rats in a maze or cats in a puzzle box might very well be different in essential respects from that of school children facing the

intricacies of a foreign language or learning to manipulate the abstract symbols of algebra. Even when human subjects have been used, as they have been in many experiments, they have usually been adults, often students of psychology. Frequently, also, the number of subjects has been small with a poorly equated control group, or with none at all. To be most convincing as a guide for school practice these experiments must be performed with children of approximately the same age and type as those found in the schools where the results are applied. If our studies of either maturation or learning are to be relied upon, the child is not merely a miniature adult. And if the numerous studies of individual differences from the time of Galton to the present have established any thing it is that what is true of one person is not necessarily true of another. Experiments based on a handful of subjects, therefore, must be accepted with considerable discount.

In the third place, the conclusions drawn have often gone far beyond the experimental facts available. This is mainly the result of the two limitations already mentioned. If experimenters would be content to draw conclusions from, and to make applications to, the same or closely similar subjects and to the same or closely similar tasks, no harm would be done. But this is rarely the case. To generalize from one age level to another is risky even when the task is the same, but it is particularly hazardous when the activity itself is different. Yet this very thing is commonly done. A meaningless and often trivial act is performed by adults under the highly artificial conditions of the laboratory. Then the results are applied without qualification to the meaningful learning of children under the actual conditions of the schoolroom. That this procedure is wholly unwarranted will appear from an experiment by the writer to be reported later in the chapter. But to generalize from the behavior of a rat in a psychological laboratory to that of a child in an ordinary classroom is little less than foolhardy.

Types of motivation experiments. From what has just been said, it may appear that the writers believe all motivation experiments to be worthless. Such is far from their conviction, however. While they believe that practically all such experiments so far reported have certain weaknesses to which attention should be called, they are convinced that genuine progress has been made and that the way has been opened for further studies to supplement those already in existence. It has been definitely shown that the problem, although difficult, is susceptible to experimental attack. Furthermore, experimental evidence already available is sufficient to provide at least tentative answers to two important questions:

- 1 What is the relation of measurement to the *amount* and *quality* of learning?
- 2 What is the relation of measurement to the *type* of learning or to the *learning procedure* followed by the student?

These two questions will now be considered.

I The Relation of Measurement to the Amount and Quality of Learning

There is considerable experimental evidence regarding the influence upon the amount and quality of learning of three measurement factors, or groups of factors

- a The frequency of the tests
- b The knowledge that a final examination would be given
- c The knowledge of results or progress in learning

Attention has been given to the operation of these factors individually, in combination with each other, and in combination with other motivating factors such as praise and blame rivalry, and various types of material rewards. Some of these findings will now be summarized.

Frequency of tests. Practice varies widely regarding the frequency of testing. At one extreme are the teachers who give no written examination of any kind, and at the other extreme are those who give a test of some kind every day. What experimental evidence is there to indicate the proper frequency? Manifestly whatever advantage may exist in frequent testing cannot be attributed solely to the motivating effects, however, since the additional practice afforded by taking the extra tests must also be considered.

The experimental evidence regarding proper frequency of testing in social studies classes in high school has not been very convincing. Hoglan⁶ studied the frequency of testing in American history in one Iowa high school. He found no significant differences among three groups, equated on the bases of intelligence and knowledge of history. One group had daily tests; another had three unannounced tests per week; and a third had only the regular tests at intervals of six weeks. In a similar study in the same subject Camp⁷ found a slight (but not a statistically significant) difference between one group tested once or twice a week and another tested once in two or three weeks. In an earlier study in community civics Shore⁸ found no advantage in giving each day a true-false test of ten items, but a group given an unannounced test two or three times a week did show a statistically significant superiority over a group given only the mid semester and the final tests. On the whole, the evidence appears to favor slightly the practice of giving a test once or twice a week to classes in the social studies in high school.

Two experiments on the effects of frequent testing of high school biology students reported conflicting results. Kitch⁹ found that the group taught with the aid of self scored unit tests did significantly better than the group

⁶ Unpublished master's thesis, University of Iowa, 1932.

⁷ Unpublished master's thesis, University of Iowa, 1931.

⁸ Unpublished master's thesis, University of Iowa, 1925.

⁹ Loran V. Kitch, unpublished master's thesis, University of Southern California.

without such tests Gable¹⁰ compared the merits of three procedures. One group was told that it would be tested each day, another group that it would be given announced unit tests, and a third group understood that it would be tested without notice at irregular intervals. On the whole, the poorest record was by the group taking daily tests, but there was a tendency for the slower pupils to do better when a test was announced which gave time for review.

Conner¹¹ found that the use of a well known series of instructional tests in high school physics had not resulted in sufficient improvement in learning to justify the time expended. Kugle¹² reported that short daily tests in physics resulted in pupils' having a small superiority over those to whom tests were given only at the ends of units. Kirkpatrick¹³ found a distinct advantage in the 26 high school physics classes included in his study, in giving an objective test at the beginning of each unit. As each unit covered from one to three days, this meant that tests were given at least twice a week. The pupils had definite knowledge that the test would be given, that it would cover all the important concepts of the unit, and that the final examination would include only points included in these unit tests. The tests were corrected in class and were used as a basis for class discussion and subsequent study. Both experimental and control groups took the same term tests at intervals of six weeks. When the experimental groups were considered as a whole, a highly significant statistical difference was found on a test of objective information given at the end of the course, but this superiority had largely disappeared four months later. The testing program was most beneficial to the pupils in the lowest third in mental ability. This suggests that schools attempting to group pupils according to ability may very well consider varying the testing program as well as the curriculum and teaching methods.

Experiments involving the frequency of testing have been most numerous and, on the whole, most convincing on the college level. A serious limitation, however, is that they have been largely restricted to classes in general and educational psychology. Jones,¹⁴ in a pioneer study, gave five-minute completion tests euphemistically called "terminal reviews," at the end of each of 27 lectures in psychology. Eight weeks later the groups so tested made scores on a final examination that were approximately twice as high as those of groups who had had no "terminal reviews." Another

¹⁰ Sister Felicita Gable. *The Effect of Two Contrasting Forms of Testing Upon Learning*. Baltimore: Johns Hopkins Press, 1936.

¹¹ Unpublished master's thesis. University of Iowa, 1932.

¹² Unpublished master's thesis. Pennsylvania State College, 1936.

¹³ James Earl Kirkpatrick. "The Motivating Effect of a Specific Type of Testing Program." *University of Iowa Studies in Education* 9: 41-68, June 15, 1934.

¹⁴ Harold E. Jones. "Experimental Studies of College Teaching." *Archives of Psychology* 68: 36-70, November 1923.

study¹⁵ reports advantages to be gained in using weekly objective tests in general psychology

Both Turney¹⁶ and Keys¹⁷ found weekly tests in educational psychology better than tests given less frequently. Turney found that when given weekly tests a class which was well below another class at the beginning was able to equal the achievement of the other class which had only one short test, in addition to the mid semester and the final, which both groups had. Keys found that eight weekly tests gave an advantage over the same items given to equivalent groups in the form of two monthly tests. However, on an unannounced examination covering the same material given five weeks later, this advantage had been reduced. When the regular final examination came after the additional two weeks the achievement of the experimental and control groups was practically identical. What the effect of the weekly testing was after still larger intervals is unfortunately unknown.

Johnson¹⁸ compared the effect of written unit tests and the effect of an equal amount of time devoted to oral reviews with 55 pairs of freshman girls in two classes in child development. She found that a statistically significant difference in favor of the tested group had disappeared twelve weeks later. She concluded that "there is as yet no evidence to show that the greater achievement which has been induced by examinations persists after six weeks to three months."

A few studies have reported little or no advantage in weekly tests even when comparisons were made at the end of the course. For example, weekly tests in general psychology at the University of Minnesota gave negative results.¹⁹ Both Noll²⁰ and Ross and Henry²¹ found a slight superiority in less frequently tested groups in educational psychology. However, Ross and Henry in both general and educational psychology, and Noll in educational psychology, found evidence that the benefit of weekly tests was greatest for the students of low ability. It is evident that there is no one

¹⁵ C. C. Ross and Lyle K. Henry. The Relation between Frequency of Testing and Progress in Learning Psychology. *Journal of Educational Psychology* 30: 604-611. November 1939.

¹⁶ Austin H. Turney. The Effect of Frequent Short Objective Tests upon the Achievement of College Students in Educational Psychology. *School and Society* 33: 760-762. June 6, 1931.

¹⁷ Noel Keys. The Influence on Learning and Retention of Weekly as Opposed to Monthly Tests. *Journal of Educational Psychology* 25: 427-436. September 1934.

¹⁸ Bess E. Johnson. The Effect of Written Examinations on Learning and on Retention of Learning. *Journal of Experimental Education* 7: 55-62. September 1938.

¹⁹ A. C. Eurich, H. P. Longstaff, and M. Wilder. *The Effective College Curriculum as Revealed by Examinations*, pages 333-347. Minneapolis: University of Minnesota Press, 1937.

²⁰ Victor H. Noll. The Effect of Written Tests upon Achievement in College Classes: An Experiment and a Summary of Evidence. *Journal of Educational Research* 32: 345-358. January 1939.

²¹ C. C. Ross and Lyle K. Henry. *op cit*, pages 609-610.

best testing technique which is equally effective under all conditions. Testing methods as well as other teaching procedures must consider the ability of the student as well as the nature of the subject.

Kulp²² gave the students in a graduate class in educational sociology who were below the median on the mid semester examination a weekly ten minute objective test for the next seven weeks. The students above the median were excused from these short tests. On the final examination "identical in all respects with the seven weekly tests" the superiority of the upper half was reduced considerably, probably due largely to practice and regression effects rather than to increased motivation. Pressey²³ reports an interesting variation of this procedure as used by Smeltzer in educational psychology. Both experimental and control classes were given weekly tests. But in the experimental class to whom the test was given on Thursday of each week, the papers were returned and discussed on Friday. Those who had made unsatisfactory scores were tested again over the same material after a brief review on Monday while the others were excused. On the final examination the experimental group was above the control group the advantage being largely with the pupils who were in the lowest fourth of the class and who had taken the retests.

Three of the above experiments attempted to get the students' attitude toward the frequent testing. By means of unsigned questionnaires in three classes Jones found that 70 per cent of the students approved the "terminal review method." In like manner Turney discovered an "excellent attitude" toward frequent testing in his experimental group about 85 per cent thought they had studied more and over 90 per cent said that they preferred to be in that section and that they felt they had learned more even if they had made no better grade. From "an extensive questionnaire touching some thirty issues of educational theory and practice," given at the opening and repeated near the end of the semester, Keys found "Without comment by the instructor or knowledge of the experiment in progress students disclose a strong and growing conviction of the desirability of tests given as frequently as every second third or fourth class session." The evidence strongly suggests that students favor frequent testing.

Awareness of final examination. To what extent is the "intention to remember" or 'temporal set' a factor in learning? Will the expectation that the material will have to be recalled later influence the amount retained? More specifically how will the awareness of a final examination affect the progress of learning and of forgetting? One or the other of the following 'two rival and mutually exclusive hypotheses' as suggested by Remmers²⁴ is apparently true

²² Daniel H. Kulp II. Weekly Tests for Graduate Students? *School and Society* 38: 157-159, July 29, 1933.

²³ Sidney L. Pressey. *Psychology and the New Education*, pages 363-366. New York: Harper & Brothers, 1933.

²⁴ H. H. Remmers and others. Exemption from College Semester Examinations. *University Studies in Education* 11: November 1933.

1. Exemption from final examinations with its requirement of continuous high level learning provides better motivation and therefore more permanent learning and integration than does the final examination

2. The final examination provides the opportunity and at least a part of the stimulation for the better development of certain abilities such as rapid organization of a large mass of material, the ability to select crucial data from the large mass of material to see pertinent relationships, to reason in terms of the subject matter, to apply this reasoning to significant problems etc. and in general more effective and permanent learning

Thisted and Remmers¹ summarized the literature on the general problem, including studies on such dissimilar materials as stories, objects shown, vocabulary, nonsense syllables, photographs, and stylus maze. They concluded "It is evident that a condition of expectation of recall when injected into the initial instructions has given variable and conflicting results." Their own study, which included 404 psychology students involved learning Anglo-Saxon vocabulary and the factual content of two articles presented in mimeographed form under ordinary class-room conditions. The control group understood that they were to be tested immediately, and the experimental groups understood that they were to be tested later also after three days in some cases, after one week and after two weeks in other cases. The experiments tended to establish a somewhat slower drop in the forgetting curve when a set to prepare for delayed recall was introduced.

While the learning material in the above experiments was not left in the hands of the students, it is reasonably sure that one effect of the "temporal set" was to cause those who expected to have to recall the material later to give a "mental review" of what they could remember, as well as probably to exchange ideas with other students. Under ordinary school conditions one might expect that the effect of reviewing for an expected examination might be larger. However, Remmers² later found that exempting students in mathematics and applied mechanics made relatively little difference in the amount, quality, or permanence of learning at least as measured by current types of tests and examinations.³

Pease⁴ reported some interesting studies of the effect of cramming on the amount of class materials retained. The first study included several classes—in all, 302 college students and 106 high school pupils—separated into equivalent groups on the basis of intelligence. A set of 100 objective items was prepared for each class, 'covering several months of the usual course work already completed by the students.' At a meeting of the class the purpose of the experiment was clearly explained. The experimental group in each class was then dismissed with instructions to spend at least an hour in review for the examination which was to come at the next meeting of the class. The control group in each class took the examination

¹ W. N. Thisted and H. H. Remmers, "The Effect of Temporal Set on Learning," *Journal of Applied Psychology* 16: 257-268, June 1932.

² H. H. Remmers and others, *op cit*, page 52.

³ Glenn R. Pease, "Should Teachers Give Warning of Tests and Examinations?" *Journal of Educational Psychology* 21: 273-277, April 1930.

at once without warning or review. The mean score of the experimental group exceeded that of the control group in each class, the average superiority being 11.1 points on the 100-item test. Without warning the test was repeated six weeks later. The average lead of the experimental group had been reduced to 6.3 points. But there was still a significant difference for all classes containing as many as fifteen pairs of pupils. However, when one of these classes was retested after an additional six weeks, the lead of the experimental group was reduced by about half and was not then significant. After twelve weeks the lead in another class had been reduced from 17.07 points to 2.7 points which was not a significant difference. These results had been produced by an amount of "cramming" by the experimental groups that represented about one and one half hours, on the average. It appeared probable that the time so spent yielded returns that averaged higher than the same amount of time spent either in class attendance or in regular preparation outside. Pease concludes that "from the standpoint of the student, it pays to cram."

Tyler and Chalmers²³ studied the effect on test results of warning junior-high school pupils that they would have a unit test in general science on the following day. The test scores of pupils so warned were compared with comparable pupils who had no specific warning although they were all aware that it was customary to have a test at the end of each unit, usually with the time announced at least two days in advance. All of the obtained differences favored the warned groups but by margins below the level of statistical significance. Six weeks later when the tests were repeated, the differences had practically disappeared. The authors questioned whether junior high school pupils are really motivated to study for unit tests even when announced, or know how to study effectively when they try. To be effective motivation has to be intelligently directed.

White²⁴ conducted an experiment that bears directly upon the effect of exemption from a final examination. Three classes in general psychology which met once a week for seventeen weeks were divided, according to chance into experimental and control groups. At each weekly class meeting both groups were given a "comprehensive mimeographed true-false test covering the chapters studied for the period." From the outset the control groups understood that their marks in the course would be based solely upon these weekly tests while the experimental groups understood that they were to have a final examination that would count 50 per cent toward their course marks. At the class meeting following each test the corrected papers were returned to all students, and they were allowed to keep the papers. The experimental groups were urged to preserve these test papers.

²³ T. T. Tyler and T. M. Chalmers. The Effect on Scores of Warning Junior High School Pupils of Coming Tests. *Journal of Educational Research* 37: 210-236, December 1943.

²⁴ Hubert B. White. Testing as an Aid to Learning. *Educational Administration and Supervision* 18: 41-46, January 1942.

for further study, as the final examination would contain exactly the same items. At the end of the seventeen weeks the final examination was given to both groups, the hearty co-operation of the students in the control groups being asked in order to determine the value of the experiment. The difference between the groups was 51.2 per cent, the experimental group having gained 31.6 per cent and the control group having lost 19.6 per cent. Even more convincing was the equal superiority of the experimental group on a completion test "with which they were wholly unfamiliar."

Knowledge of test scores. What is the effect upon the course of learning of the knowledge of progress, afforded by test scores or by other means? The answer to this question has been sought many times in the psychological laboratory, with practically unanimous results. Psychologists are in substantial agreement with the conclusion of an early study²⁰ that "the addition of a single other motivating factor, namely, knowledge of results, is sufficient to give the pupils with such knowledge a distinct superiority over the others, and the degree of superiority is roughly proportional to the amount of information possessed." However, as has been pointed out earlier in the chapter, experiments conducted in the classroom are far more convincing. We shall now take a look at what they have shown.

One of the earliest and most comprehensive of these studies conducted under actual schoolroom conditions was that of Panlasigui.²¹ The findings were based on 358 pairs of pupils in fourth-grade arithmetic in ten cities. The practice material consisted of fifteen minutes' drill in examples of the mixed type of fundamentals once a week for twenty weeks. As all pupils scored their papers after each drill, it can be seen that each pupil knew his achievement for the day, although this knowledge must be related to previous records in order to be a knowledge of progress, strictly speaking. In the experimental classes the idea of progress was stressed, progress charts for both the individual and the class being kept in a conspicuous place. The teachers of the control classes, on the other hand, were instructed as follows: "Please keep very much out of class discussion any reference about how much pupils are scoring." The comparison appeared to be, then, between experimental classes with somewhat more stress on progress, and control classes with somewhat less stress on progress, than is customary to the ordinary teacher. On a comprehensive test the mean of the experimental classes exceeded that of the control classes by 11.34. A detailed examination of the results reveals the fact that this superiority is most in evidence in the highest quarter and practically non-existent in the lowest quarter. "The beneficial effect of awareness of success, then, was substantially in direct proportion to the amount of success available for motivation." This is also

²⁰ See pages 306-307 for a brief description.

²¹ Isidoro Panlasigui, *The Effect of Awareness of Success on Skill in Arithmetic* unpublished doctor's dissertation, Iowa City University of Iowa, 1928. For a brief account see *Twenty Ninth Yearbook of the National Society for the Study of Education*, pages 611-619. Bloomington, Illinois: Public School Publishing Company, 1930.

true of the drill periods themselves, where the accuracy standards of the highest fourth of the experimental group exceeded those of their controls eight times out of eleven, whereas the lowest fourth of the experimental groups fell behind their controls on every drill. This experiment seems to have established rather definitely two important points:

1. A knowledge of progress in learning under classroom conditions is likely to have much less effect than that under laboratory conditions.

2. A knowledge of progress is likely to be more beneficial to good students than to poor.

Studies since reported have confirmed the first point, but most of them have not been analyzed with respect to the second point.

Forlano³² conducted a comprehensive series of experiments in grades four to eight, inclusive, involving in all 1,294 pupils, and touching upon various aspects of the problem of the effect on learning of a knowledge of results. The experimenter emphasizes the fact that these studies were made "in the normal classroom situation as far as possible as a part of the daily school routine." He attempted to determine whether giving a knowledge of results *immediately* after the word had been spelled or an arithmetic fact had been studied was more effective than when a knowledge of results was *withheld* until an entire column of 20 or 24 items had been attempted. In other words, if one may use the analogy of the target range, Forlano was interested in finding out, so to speak, whether it was better to tell the marksman his score after each shot or to wait until he had fired a series of 20 shots. The author's conclusion is as follows:³³

The results of our experiments show that there is a tendency for learning during which the learner ostensibly receives immediate knowledge of results to be less efficient than learning in which knowledge of results is delayed. In general, it may be said that this superiority of the "delayed knowledge of results" method does not always approach statistical certainty.

Even this modest conclusion, the author suggests, is limited by the fact that the methods employed "may not be 'pure' methods of what they purport to involve," and that the apparent superiority of the delayed procedure may be due to other causes. In any event, since the period of delay never exceeded five minutes, little light is shed upon the ordinary school situation, where the tests follow learning after an interval ranging from a day to a year or longer.

Brown³⁴ reports an experiment in arithmetic in grades 5A and 7A. Both his procedure and his conclusions differed somewhat from those of Pan-

³² George Forlano, *School Learning with Various Methods of Practice and Rewards*, pages 55-114. New York: Bureau of Publications, Teachers College, Columbia University, 1936.

³³ *Ibid.*, page 99.

³⁴ Francis J. Brown, "Knowledge of Results as an Incentive in Schoolroom Practice," *Journal of Educational Psychology* 23: 532-552, October, 1932.

design. In grade 7A, Brown selected his experimental and control grades which were only roughly equivalent, on the basis of an intelligence test, and in grade 5A on the basis of estimates of intelligence and achievement. The groups were reversed at the end of the first period of ten days. The drill period was eight to ten minutes daily. While the differences on the whole, favored the experimental group, they were not very impressive. An examination of the individual drill periods reveals the fact that the progress from day to day in all groups was irregular and somewhat inconsistent, and that the differences between experimental and control groups were generally less on the tenth day than on the first. There was some evidence in Brown's study that the incentive was somewhat more effective with boys than with girls but the outstanding fact was the remarkably small amount of influence taken from any point of view, of a knowledge of progress in the classroom as compared with the laboratory.

Deputy²⁵ conducted in a state university a carefully planned experiment with three groups of students, of approximately equal intelligence in freshman philosophy, which met twice a week. For six weeks during the first half of the semester the first ten minutes of each class meeting of the control group were devoted to an oral review of the preceding lesson. One of the experimental groups had a ten minute objective test covering the same material, and the other experimental group had the same items in a twenty minute test given once a week. Beginning at the middle of the semester the group which had served as a control was given the ten minute test at each class meeting while the other two groups had only the oral reviews. The scores for the experimental groups were put on the board following each test, and each student was urged to keep a record of his progress. Only one of the three comparisons between the ten minute written test and the ten-minute oral review showed the former to be superior by a statistically significant amount. This fact the author ascribed to a particularly favorable attitude on the part of the students. The experimental group which excelled happened to be slightly the most intelligent of the three and also showed itself superior to the group which took the twenty minute test once a week. Deputy's most significant conclusion was: Considerable precaution should be taken in applying principles derived from laboratory and other non classroom situations to work in school subjects.

Two years later Ross²⁶ began a series of experiments which were to force him to this same conclusion. Attention has already been called to the earlier laboratory experiment²⁷ which had appeared convincing not only to the author at the time but also to many readers since that time judging

²⁵ E. C. Deputy, *Knowledge of Success as a Motivating Influence in College Work*, *Journal of Educational Research* 20: 327-334, December 1929.

²⁶ C. C. Ross, *The Influence upon Achievement of a Knowledge of Progress*, *Journal of Educational Psychology* 24: 609-619, November 1933.

²⁷ See footnote 5.

from the writers on educational psychology who have quoted it with approval

Upon the basis of a comprehensive examination given at the end of the first unit in a class in tests and measurements, a large class was divided into four substantially equivalent groups. A regular class test was given to all students once a week for the next two months. At the next class meeting following a test, a distribution for the entire class was put on the blackboard and a brief discussion given of each item missed by any considerable number. But the four groups were given different degrees of information as to progress. One group was given *no knowledge* whatsoever as to its scores. A second group was given *vague knowledge*, each student being told merely that his score was "good," "fair," or "poor." A third group was given *partial knowledge*, each student being told his point score but not allowed to see his paper. The fourth group was given *full knowledge*, each student being shown his paper at the close of the class and allowed to ask any questions he wished to ask regarding it.

Figure 38 shows the results for the four groups in the form of cumulative scores, week by week, for the first eight weeks and for the last four weeks,

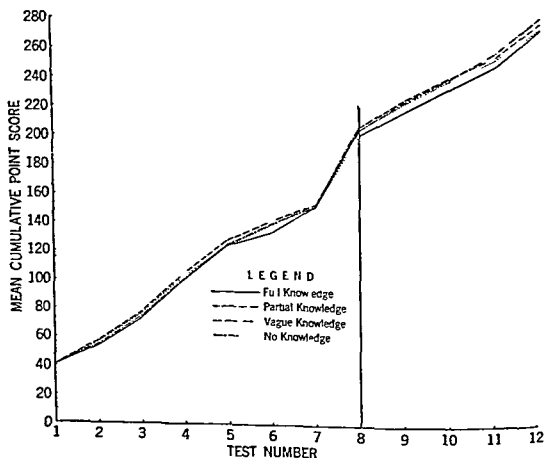


Figure 38 The Influence of Knowledge of Progress Upon Achievement in a College Class

when the groups were reversed. Nowhere was there a statistically significant difference between any of the groups. The experiment was repeated with two other classes in the same subject and with one in a different subject. Not content with this, Ross persuaded a colleague in another department to do the same experiment. In all the groups involved more than 50 tests and about 300 students and not once did there appear a difference favoring the group with full knowledge of progress that meets the minimum requirement for statistical significance.

Two conclusions seem reasonably certain. The first directly in line with that of Deputy is as follows:³⁸

The Gestalt of the laboratory situation is so different from that of the life situation outside that it is hazardous to generalize from one to the other. One can never be certain what the outcome of a laboratory experiment will be when applied to the classroom situation until it has actually been tried out in that situation.

The second conclusion is that most if not all experiments relating to knowledge of results in learning have involved another erroneous assumption, namely, that because students were not told their individual scores they "had no knowledge of progress." Certainly they had their subjective impressions. To test out the accuracy of these impressions the author requested the students in the "no knowledge" group to estimate the scores they thought they had made when they turned in their papers at the close of the tests. The median coefficient of correlation between these estimates and the actual scores was .71. Manifestly then, such studies involve a comparison of *two kinds of knowledge*, subjective and objective. Moreover, there was a tendency for the poorer students to overestimate their scores. In such cases the illusion of success may very well have proved more stimulating than the reality of failure.

Knowledge of results combined with other incentives. It is probably rare that a knowledge of progress operates alone. It is likely that such factors as rivalry and social recognition are always involved in some degree. But in the experiments so far reported, these other factors were not emphasized. In many experiments, however, the knowledge of progress has merely been taken as the *occasion* for utilizing other motives, such as praise and blame, rivalry, money or other rewards, and the like.

At least two studies have attempted to use a knowledge of intelligence scores as an occasion for verbal suggestion and other forms of motivation. Mitchell³⁹ divided the lowest fourth of the freshman class in a high school into two equivalent groups on the basis of the Otis tests. Each pupil in the experimental group received the following notice without further comment:

³⁸ C. C. Ross, "A Needed Emphasis in Psychological Research," *Psychological Review* 43, 197-206, May 1936.

³⁹ Claude Mitchell, "Why Do Pupils Fail?" *Junior-Senior High School Clearing House* 9, 172-176, November 1934.

Dear Pupil

Your score on the Intelligence Test which was given at the opening of school is LOW. This will mean that much work and effort on your part will be necessary to keep up with the class. Put yourself to the task and show that you can do it. **YOU CAN IF YOU WISH**

Principal

At the end of the year it was found that 62 per cent of the group which had received this notice passed on all subjects, while only 15 per cent of the equally poor group, which had not been notified, did so.

Ross⁴⁰ conducted a somewhat similar study with college students at the University of Kentucky. From the lowest fifth in intelligence, experimental and control groups of 40 freshmen each were formed upon the basis of psychological tests, sex, and fraternity affiliation. The students in the experimental group were then called together, and a frank statement was made regarding their scores. They were told that it was important at the outset to recognize the fact that they were up against a somewhat different situation from that of the students with higher test scores. They were as-

TABLE 43

POINT STANDING FOR THE FIRST AND SECOND SEMESTERS FOR LOW-RANKING FRESHMEN WHO WERE TOLD THEIR INTELLIGENCE TEST SCORES AS COMPARED WITH THOSE WHO WERE NOT

POINT STANDING	FIRST SEMESTER						SECOND SEMESTER					
	Arts & Sci		Commerce		Total		Arts & Sci		Commerce		Total	
	Exp *	Con †	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con
180-199			1		1							
160-179		1	1		1	1	2	3			2	3
140-159	3	1			3	1	1	1	1		2	1
120-139	4	2	1		5	2	2	2	3		5	2
100-119	7	2	2		9	2	4		1		5	
80-99	2	5	1	2	3	7	1	6	1	1	2	7
60-79	4	5	3	3	7	8	3	3	5	1	8	4
40-59	3	3	5	2	8	5	3	1	2	6	5	7
20-39		3	2	7	2	10	1	4		3	1	7
00-19	1	2		2	1	4	3	1	1	2	4	3
Total	24	24	16	16	40	40	20	21	14	13	34	34
Mean	98	78	83	45	94	64	85	88	86	44	85	69
S D	36	41	46	25	41	39	50	49	45	22	45	46
$M_E - M_C$	20		38		30		- 03		42		16	

* Experimental group

† Control group

⁴⁰ C. C. Ross, "Should Low-Ranking College Freshmen Be Told Their Scores on Intelligence Tests?" *School and Society* 47: 678-680, May 21, 1938.

sured, however, that the experience at the University showed that such students could succeed, if they were willing to work and did not attempt too heavy a load in school or too many activities outside. The control group had no advance information.

The record of the two groups is summarized in Table 33. The mean point standing of the experimental group was 94 for the first semester and 85 for the second semester, while the corresponding values for the control group were 64 and 69 respectively. During the first semester three times as many students in the experimental group as in the control group made a point standing of 100 or better, and more than twice as many made this standing the second semester. Approximately twice as many experimental as control students passed all subjects. On the whole the difference was more marked for the first semester than for the second, and was decidedly greater for the College of Commerce than for the College of Arts and Sciences. These two studies offer rather convincing evidence that knowledge of intelligence test results may have a motivating effect on low-ranking freshmen in high school and college. More recent studies have tended to confirm these findings.⁴¹

A great many more studies have utilized achievement test scores as occasions for various types of motivation. In an early study Book and Norvell⁴² used a knowledge of results in four laboratory experiments as a basis for building morale or developing the "will to learn." For example, students in the experimental groups "were frequently told that if they would only make up their minds to increase their score they would somehow find a way to do it," while at the same time the "method of measuring their output and having them keep track of their score usually convinced them that this was true." Their data support the conclusion that this "special group of incentives" help the experimental group to "make more improvement with a given amount of practice than do the control groups." But it is impossible to tell just how important a knowledge of results by itself would have been. An experiment by Hurlock,⁴³ which utilized test results as occasions for praise and reproach, attracted considerable attention. The subjects were 106 children in fourth- and sixth grade arithmetic. The groups were equated on an initial practice period of fifteen minutes. Four more practice periods were held on successive days. The control group received the tests without comment. The praised group had their names read aloud at the beginning of each practice period. They were then called to the front of the room and

⁴¹ Cf. R. K. Compton, "Student Evaluation of Knowing College Aptitude Test Score," *Journal of Educational Psychology* 32: 656-664, December 1941.

Edna F. Iampson, "How Objective Can Freshmen in College Be toward Objective Evidence of Their Ability and Achievement?" *Educational Administration and Supervision* 28: 280-290, April 1942.

⁴² William F. Book and Lee Norvell, "The Will to Learn: An Experimental Study of Incentives in Learning," *Pedagogical Seminary* 29: 305-362, December 1922.

⁴³ Elizabeth B. Hurlock, "An Evaluation of Certain Incentives Used in School Work," *Journal of Educational Psychology* 16: 149-159, March 1925.

received praise combined with exhortation to do still better work. Then the names of the children in the reprovod group were called, and they were severely reprovod for poor work, carelessness, and general inferiority. The ignored group heard what was said to the others, but they received no recognition whatsoever. The results are shown in Figure 39. After the first

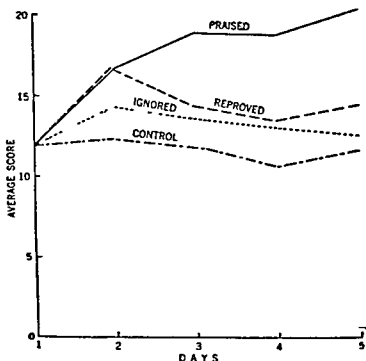


Figure 39. A Study of the Influence of Praise and Reproof upon Achievement in Fourth Grade and Sixth-Grade Arithmetic (After Hurlock)

day reproof seemed far less effective than praise, although somewhat better than being ignored altogether. The control group made no progress whatsoever. It is to be regretted that this experiment was not continued for several days longer. Manifestly an hour's total working time is insufficient to establish fully the comparative merits of these incentives as they would operate day after day in the ordinary classroom.

In a somewhat similar experiment in the same grades, Hurlock⁴⁴ studied the effect of group rivalry on addition. The control group took their tests for ten minutes on four days without comment. The experimental group was divided into two equivalent subgroups which were pitted against each other. The author emphasized the fact every day that the two groups "were absolutely equal, and that one had as much chance to win as the other." Although the effect of rivalry was present in all types of pupils, it was most marked in younger pupils and in inferior pupils. Increase in accuracy was much less than increase in speed, with some tendency for increase in speed.

⁴⁴Elizabeth B. Hurlock, "The Use of Group Rivalry as an Incentive," *Journal of Abnormal and Social Psychology* 22: 278-290, October-December, 1927.

to be accompanied by reduction of accuracy. It is well to keep in mind Thorndike's warning that "the attainment of active rather than passive learning at the cost of practice in error may often be a bad bargain."⁴⁵

Another study⁴⁶ shows that repeated applications of praise or blame may have different effects on introverted and extroverted pupils. Introverted fifth grade pupils improved faster in number cancellation exercises when praised than did either introverts who were blamed or extroverts who were praised. However, extroverted pupils when blamed improved faster than extroverts who were praised or introverts who were blamed. Unfortunately one cannot safely conclude from a study which involved a total practice time of three minutes upon highly artificial tasks that the same differences would necessarily appear under ordinary school conditions. The problem is worthy of further experimentation.

II The Relation of Measurement to the Type of Learning

Closely related to the amount and quality of learning is the type of learning or the learning procedure which is employed. There is considerable evidence for thinking that effective work or study habits of the student are of fundamental importance in learning. A question of major importance therefore, is to what extent does the type of measurement used influence the type of study technique employed by the student? Some important studies bearing on this question have been conducted on the college level.

In a pioneer study, Terry⁴⁷ found that 236 students in educational psychology were "influenced to a significant extent by the type of examination for which they were preparing. The most striking characteristic of the methods employed in preparing for an objective test which had been announced a month in advance was the students' emphasis on details while they tended to study for large units of subject matter when they were preparing for an essay examination announced for the next month. Douglass and Tallmadge⁴⁸ reported similar results at the University of Minnesota. They found that the 'objective type focuses attention upon details and exact wording while the subjective type apparently favors methods involving organization, perceiving relationships and trends and personal reactions'."

There also appear to be significant differences among the various forms of the so-called new type examinations in their effect on study methods

⁴⁵ Edward L. Thorndike and others *The Psychology of Wants, Interests, and Attitudes* New York: D. Appleton Century Company, 1935, Page 147.

⁴⁶ George G. Thompson and Clarence W. Hunnicutt, "The Effect of Repeated Praise or Blame on the Work Achievement of Introverts and Extroverts," *Journal of Educational Psychology* 35, 257-266, May 1944.

⁴⁷ Paul W. Terry, "How Students Review for Objective and Essay Tests," *Elementary School Journal* 33, 592-603, April 1933.

⁴⁸ Earl R. Douglass and Margaret Tallmadge, "How University Students Prepare for New Types of Examinations," *School and Society* 39, 318-320, March 10, 1931.

Terry⁴⁹ found for example, that the one predominant method of preparing for completion tests emphasized the word for word mastery of statements considered important while preparing for true-false tests involved methods which dealt primarily with definitions and detailed facts such as the authors and findings of experiments. The author's conclusion points out an important educational implication.

The kind of test to be given if the students know it in advance determines in large measure both what and how they study. The behavior of students in this habitual way places greater powers in the teacher's hands than many realize. By the selection of suitable types of tests the teacher can cause large numbers of his students to study to a considerable extent at least in the ways he deems best for a given unit of subject matter.

Meyer⁵⁰ conducted a careful laboratory experiment with 124 psychology students to determine the relation between the specific examination set and immediate memory and delayed memory after five weeks. When the *amount* of study was held constant the *method* and *results* appeared to be largely dependent on whether the set was for recall or for recognition tests. It appeared that when students expected completion tests they studied with more effort than they would have put forth for recognition tests. More students made summaries and maps and otherwise attempted to obtain a general picture of the material when they expected essay examinations than otherwise. Meyer points out four practical implications.

1 Since it is more economical when a given amount of time is spent in studying to use a recall examination set for delayed recognition or immediate and delayed recall tests recognition questions should be used in testing only when they form a part of the entire examination or when students are unaware that such questions are to be used exclusively.

2 If the teacher feels it necessary that the students be able to recognize certain materials for a short time only then the indications are that a recognition examination set may be used. This means that the teacher must evaluate the material in his course very carefully since recognition tests if given indiscriminately may have a deleterious effect on what the students ultimately retain of the course.

3 If the teacher feels it necessary that the students be able to recall isolated facts when specific cues are given as to the fact wanted a completion examination set may be used with profit.

4 If the teacher wants the students to recall the material in an organized fashion and to know facts when cues are not given the essay examination set should be used in preference to any objective type of examination set. Here again the teacher must evaluate the material which he presents in the light of what the student should learn from the course.

⁴⁹ Paul W. Terry. How Students Study for Three Types of Objective Tests. *Journal of Educational Research* 27 333-343 January 1934.

⁵⁰ George Meyer. An Experimental Study of the Old and New Types of Examination. *Journal of Educational Psychology* 25 641-661 December 1934 and 26 30-40 January 1935.

The following quotation from Monroe⁴¹ suggests that the nature of the examinations emphasized by the teachers may influence the students' reactions much more than the objectives of the course

There has been much discussion of the importance of teachers formulating their objectives and, in response to the pressure of authority, they have spent many hours in formulating lists of immediate objectives, that is the goals toward which students are expected to direct their efforts. Many of these lists merit commendation, but their influence upon students is practically nil in comparison with the influence of the tests administered. Students direct their efforts toward becoming able to respond to the tests they anticipate.

D. Some Educational Implications of Motivation Studies

Much of the experimental evidence on motivation has been fragmentary, some of it contradictory, and hardly any of it conclusive. But a few generalizations appear to have been fairly well established.

Implications for educational theory. In the first place, there is grave danger of premature and unwarranted generalizations in psychology and education. That it is hazardous to generalize from the laboratory experiment to the classroom application has been demonstrated in motivation experiments again and again. It is also dangerous to generalize from one age level to another. This is one of the greatest limitations of much of the experimental work on motivation. There is a great need for comparing the results of experiments made on the college level with results obtained from pupils on the elementary and secondary school levels.

In the second place, there are no fixed motivating categories such as knowledge of results, praise and blame, rewards and punishments et cetera. Brenner states this point well.⁴²

The truth seems to be that there do not exist such psychological entities but that they do act in *specific situations depending upon all the factors of the situation as a whole*. What in one situation may constitute praise under certain other circumstances will be considered blame. The incentives derive their attributes, so to speak, from the situation in which they are active.

Implications for educational practice. Three points require brief mention. In the first place the measurement program of the school influences both the teacher and the learner. It affects teaching emphasis and curriculum content as well as the amount and quality of learning and the procedure employed. In the second place, no motivating factor operates

⁴¹ Walter S. Monroe, *Some Trends in Educational Measurement*, Twenty-Fourth Annual Conference on Educational Measurements, page 32, Bulletin of the School of Education, Indiana University, Vol. XIII, No. 4, Bloomington, Indiana: Bureau of Cooperative Research, 1937.

⁴² Benjamin Brenner, *Effect of Immediate and Delayed Praise and Blame upon Learning and Recall*, pages 48-49, New York: Bureau of Publications, Teachers College, Columbia University, 1934.

universally Both Chase and Hurlock, for example, found young children more susceptible than older children to the motivation used In general, praise seems more effective with the duller and socially inferior groups Frequent testing also seems most helpful to weaker pupils On the other hand, there is some evidence that blame and knowledge of results are more effective in the stronger groups Even in similar age and social groups, however, marked individual differences appear as to the relative effectiveness of different types of motives, or even as to the effectiveness of the same motive used at different times Brenner⁵³ warns against a

stereotyped habit of motivation for instance always praising the children always smiling and appearing pleased This form of mechanized motivation is not adequate for increasing the performance of children, and it is doubtless harmful in its influence upon character building in children

In the third place, no motivating factor operates automatically Test scores, at best, merely provide an occasion for praise or blame, reward or punishment, or some form of social recognition The strategic place of the teacher is nowhere more in evidence than in motivation In a fundamental sense, the role of the teacher is to stimulate and guide the learning process Perhaps Brenner's concluding statement⁵⁴ does not put the matter too strongly

The facts about the usefulness of a motive in a certain learning situation will be furnished by educational psychology but proper application of the incentive in a given situation depends upon the insight of the teacher The effectiveness or worth of a teacher depends upon his ability to make adequate use of motivation

E Practice Effect

The whole question of practice is intimately related to learning in general One special aspect, practice effect on repeated tests, has received considerable attention Several standardized tests, such as the American Council on Education Psychological Examination,⁵⁵ contain short pretests which help the examinee 'warm up' and acquire the proper set for the subtest that follows Some examiners prefer to preface a testing session with easy practice material to 'cushion' inexperienced testees and thereby put them more nearly on an equal footing with "test-wise" students⁵⁶ Even the effects of coaching on highly similar material may be overestimated, however as Dyer⁵⁷ has demonstrated quite well with preparatory school students

⁵³ *Ibid* page 50

⁵⁴ *Ibid* page 50

⁵⁵ Both high school and college-freshman forms are published by the Cooperative Test Division of Educational Testing Service

⁵⁶ For a surprising by product of this procedure see Scarvia B Anderson "Prediction and Practice Tests at the College Level" *Journal of Applied Psychology* 37 256-259 August 1953

⁵⁷ Henry S Dyer Does Coaching Help? *College Board Review* No 19 331 335 February 1953

Nevertheless, it is undoubtedly true that the individual who takes a standardized examination for the first time in competition with experienced examinees is handicapped, especially if the test involves speed, complexity, and novelty

SELECTED REFERENCES FOR FURTHER READING

- Cane, V R, and Heim, Alice W 'The Effects of Repeated Retesting III Further Experiments and General Conclusions' *Quarterly Journal of Experimental Psychology* 2 182-197, November 1950
- Cook, Walter W, "The Functions of Measurement in the Facilitation of Learning" Chapter 1 in E F Lindquist (Editor) *Educational Measurement* Washington D C American Council on Education 1951
- Current Theory and Research in Motivation—a Symposium* Lincoln University of Nebraska Press, 1953 193 pages Articles and comments by Judson S Brown Harry F Harlow, O Hobart Mowrer, Theodore M Newcomb, Vincent Nowlis, and Leo J Postman
- Hilgard Ernest R, *Theories of Learning* New York Appleton Century Crofts, Inc, 1948 409 pages
- Hilgard, Ernest R, and Russell David H, "Motivation in School Learning," Chapter II in "Learning and Instruction" *Forty Ninth Yearbook of the National Society for the Study of Education, Part I* Chicago University of Chicago Press, 1950
- MacKinnon, Donald W, 'Fact and Fancy in Personality Research,' *American Psychologist*, 8 138 146 April 1953
- McClelland, David C, 'The Measurement of Human Motivation An Experimental Approach,' pages 41 56 in *Proceedings of the 1952 Invitational Conference on Testing Problems* Princeton New Jersey Educational Testing Service, 1953
- McGeoch, John A, and Iron Arthur I *The Psychology of Human Learning* (Second Edition) New York Longmans, Green and Company, 1952 Chapter VI, "Learning as a Function of Motive-Incentive Conditions"
- National Council of Teachers of Mathematics *The Learning of Mathematics, Its Theory and Practice*, Twenty First Yearbook Washington, D C The Council, 1953 Chapter II by Maurice L Hartung 'Motivation for Education in Mathematics,' and Chapter VI by Ben A Suelz, 'Drill—Practice—Recurring Experience'
- Peel, E A, "A Note on Practice Effects in Intelligence Tests,' *British Journal of Educational Psychology* 21 122-125, June 1951
- Peel, E A, 'Practice Effects between Three Consecutive Tests of Intelligence,' *British Journal of Educational Psychology* 22 196-199, November, 1952
- Pressey, S L, "Development and Appraisal of Devices Providing Immediate Automatic Scoring of Objective Tests and Concomitant Self Instruction,' *Journal of Psychology*, 29 417-447, April 1950
- Tyler, Ralph W, 'The Functions of Measurement in Improving Instruction,' Chapter 2 in E F Lindquist (Editor), *Educational Measurement* Washington, D C American Council on Education 1951

12

Diagnosis

A. The Problem of Diagnosis in Education

The nature of educational diagnosis. Educational diagnosis seeks to determine the nature and causes of unsatisfactory adjustment to the school situation. It is concerned with the specific weaknesses of individual pupils. Diagnosis seeks not so much to describe or explain educational maladjustment as to correct or prevent it. Adequate diagnosis is the basis of intelligent guidance and of effective teaching.

Education borrowed the term "diagnosis" from medicine, where its fundamental character has been long recognized. Medical diagnosis commonly starts with some bodily symptom, such as pain or abnormal temperature. The next step is to determine the causes that lie behind the symptoms. The trouble may be the malfunctioning of some organ or gland, which in turn may be caused by some particular germ or toxic condition, and which, when located, may yield readily to the appropriate medical treatment or surgery. The order of events is clearly indicated by the rule "Before you dose, diagnose."

The situation in education is much the same, although here the scope of diagnosis is usually broader. At times educational difficulties can be traced to some organic defect, such as imperfect vision or hearing, or some glandular disorder, but educational diagnosis is more often concerned with functional disorders rather than organic. Pupils who are perfectly normal organically may experience great difficulty with various aspects of the school situation. It is a matter of common knowledge that many serious learning difficulties arise, not so much from structural defects as from other factors, such as faulty habit-formation, lack of interest, or a poor home environment. Despite these complications an outstanding educator has as-

serted that "experts in reading arithmetic, and spelling can now make diagnoses no less valid and reliable than are most diagnoses in medicine"¹

Furthermore, the learning process at any time is usually conditioned by many factors, both inside and outside the learner. It is rarely possible to isolate a single causative factor analogous to the disease germ in medicine but the various factors may be classified roughly as follows

1 Internal factors

- a Physical sensory equipment glandular balance health status stage of maturity level etc
- b Intellectual general intelligence specific talents and deficiencies etc
- c Emotional attitudes interests drives prejudices feelings of inadequacy etc
- d Educational background work habits etc

2 External factors

- a School environment educational program teacher playmates equipment etc
- b Extra-school environment home community church recreational facilities, etc

The scope of educational diagnosis has also increased to keep pace with the growing concept of education. When the conventional school conceived of its function rather narrowly in terms of certain academic knowledge and skills, the scope of diagnosis was likewise limited. Now that the modern school has enlarged the concept of education to make it synonymous with the growth of personality it is no longer proper to limit the scope of diagnosis to locating the causes that interfere with the ordinary academic progress of the pupil. The learning difficulties presented by the school curriculum will doubtless always constitute an important part of any program of diagnosis. In fact this phase of diagnosis naturally increases in scope and importance as the objectives of the various school subjects are extended to include the less tangible outcomes, such as attitudes interests appreciations, tastes and standards of judgment. But some of the most important and difficult aspects of diagnosis have to do with social adjustments and personality disorders of many kinds.

It is likewise apparent that the scope of diagnosis is much larger than the use of tests and examinations. This of course does not mean that tests have an unimportant place in educational diagnosis. On the contrary an adequate diagnosis may involve the use of intelligence tests both general and specific, and of diagnostic achievement tests both standardized and teacher-made, as well as the use of various pieces of laboratory apparatus for measuring sensory acuity co-ordination, and the like. In addition to many kinds of tests, reliance must be placed upon other forms of appraisal such as rating scales uncontrolled observation questionnaires and interviews. Important as are the ordinary forms of measurement in diagnosis

¹ William A. Brownell *Quantitative Research on Learning and Teaching* *School and Society* 50 851 December 30 1939

they are often by themselves insufficient. Keys has well stated the role of intelligence tests in diagnosis:²

Few psychologists today look to an individual's score on an intelligence test alone and of itself, to determine the source of his difficulties or indicate the exact solution to his problems. It is entirely probable, however, that the outcome of such a test judiciously chosen and competently administered, will contribute as much if not more to sound clinical appraisal than any other single fact obtainable. Properly supplemented with other diagnostic procedures, the information thus derived is virtually indispensable to intelligent attack upon a wide variety of problems.

The importance of educational guidance in the modern school arises from two facts: (1) many pupils make unsatisfactory progress in school—some fail altogether and others achieve little, and (2) few causes of maladjustment lie on the surface or are self-evident.

It should be noted that up to the present time most tests designed specifically for diagnostic purposes have been for the elementary school. As long as the secondary school and college had highly selected student bodies, their need for diagnostic tools was less acute. In recent years the enlarged enrollments at these higher levels of education have greatly increased the need for diagnosis.

The value of diagnosis in education. There is an abundance of experimental evidence to show the value of educational diagnosis combined with the appropriate remedial measures. Such evidence is available on all levels of instruction and in a variety of subjects. Science has added confirmation to the verdict of common sense: it really helps to "put the oil where the squeak is." For example, Baker³ found that four months' special coaching of sixty nine-year-old pupils from seven Detroit schools resulted in a gain of about seven months in educational age. The coaching consisted of two thirty minute periods per week devoted to the subject or subjects in which the pupil had shown weaknesses. Scruggs⁴ compared the improvement of two equivalent classes of fifth grade Negro children in Kansas City, one of which had the ordinary group instruction in handwriting and the other an equal amount of corrective practice based upon a detailed analysis of the weaknesses of each pupil. In seven weeks the second group increased the average quality of its handwriting about twice as much as the first. In a similar study Guiler⁵ found that fourteen seventh grade pupils made

² Noel Keys, *Applications of Intelligence Testing*, *Review of Educational Research* 8:256 June 1938.

³ Harry J. Baker, *Educational Disability and Case Studies in Remedial Teaching*, page 63, Bloomington, Illinois: Public School Publishing Company, 1929.

⁴ Sherman D. Scruggs, 'Remedial Teaching for Improvement in Handwriting', *Journal of Educational Research* 23:288-293 April 1931.

⁵ Walter Scribner Guiler, 'Improving Handwriting Ability', *Elementary School Journal* 30:56-62 September, 1929.

in three months a normal gain of three years in quality of handwriting Blair⁶ has summarized studies in the tool subjects on the secondary level which show similar results

It has been shown that the value of such remedial measures is by no means confined to skill subjects such as handwriting and spelling. For example, a study by Leonard⁷ showed that junior high school pupils improved more rapidly in the ability to write compositions free from common errors in capitalization and punctuation during a program involving error analysis and appropriate remedial exercises than did pupils of like ability exposed to the conventional method of teaching. While both groups showed definite improvement the mean decrease in the twenty-eight most frequent errors, after eleven forty-five-minute practice periods, was approximately twice as great for the experimental as for the control groups. Experiments by Guiler⁸ on the elementary school, the senior high school and the college levels showed comparable results from similar methods.

Stone⁹ found that pupils in the fifth and sixth grades in twenty-three schools, who devoted not more than forty minutes a day for five weeks to diagnostic and practice tests, gained two to six times as much in ability to solve reasoning problems as did pupils of equal ability who had only the regular arithmetic work in school. Furthermore, the results of the study indicated that the superior gain in reasoning ability resulting from this diagnostic and remedial program was about twice as great for pupils in the highest sixth in intelligence as for those in the lowest sixth; that the gain transferred to problems of a different content; and that it persisted for at least a year, at the end of which the retests were given.

The psychology of such procedures seems reasonably clear. It is a sound principle of teaching which holds that learning always begins where the learner's present knowledge leaves off. Failure to observe this principle results in foolish attempts to do two impossible things. One of these is attempting to teach a pupil what he already knows. The other is attempting to teach him on a level too far beyond his present knowledge. Both are equally futile. The only adequate safeguard to be obtained is in frequent check-ups on the pupil's progress.

⁶ Glenn Myers Blair, *Diagnostic and Remedial Teaching in Secondary Schools*, 422 pages, New York: The Macmillan Company, 1946.

⁷ J. Paul Leonard, "The Use of Practice Exercises in Teaching Capitalization and Punctuation," *Journal of Educational Research*, 21, 186-190, March 1930.

⁸ Walter S. Guiler, "Improving Instruction in English Mechanics in the Elementary School," *Elementary School Journal*, 34, 427-437, February 1934. "Improvement and Permanency of Learning Resulting from Remedial Instruction," *School Review*, 41, 430-458, June 1933, and "Remediation of College Freshmen in Sentence Structure," *Journal of Educational Research*, 26, 110-115, October 1932.

⁹ C. W. Stone, "An Experimental Study in Improving Ability to Reason in Arithmetic," *Twenty-Ninth Yearbook of the National Society for the Study of Education*, parts 1-3, 389-399, Bloomington, Illinois: Public School Publishing Company, 1930.

B. The Techniques of Diagnosis

The levels of diagnosis. The process of educational diagnosis may be profitably thought of as falling into five steps, or levels. Figure 40 is a graphical representation of the process. It will be noted from the questions asked at each level that the first four steps—the W's—have to do with

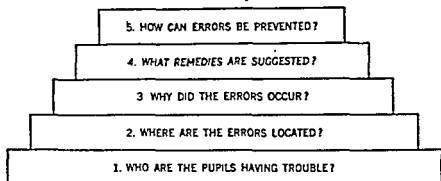


Figure 40. The Five Levels of Educational Diagnosis.

corrective diagnosis, while the highest level has to do with what may be termed preventive diagnosis. In other words, the immediate purpose is *correction*, but the ultimate purpose is *prevention*.

Locating the individuals needing diagnosis. How can we best locate the pupils not making satisfactory adjustment to the school situation? This is logically the problem with which the program of educational diagnosis begins. The order of events is not unlike that described in the famous recipe for making rabbit stew which begins: "First you catch your rabbit." Strictly speaking, however, while it is a necessary preliminary step, it is hardly a part of the actual process of diagnosis.

Various ways of locating the individuals who require diagnostic study have been used. Survey and group intelligence tests are often employed to screen those whose achievement is unsatisfactory. Using this method Wilson¹⁰ found that about 70 per cent of the pupils in the seventh and eighth grades of fifteen representative cities and towns in the metropolitan area of Boston needed corrective instruction in the fundamental arithmetic processes.

Several writers suggest that any pupil whose level of achievement is well below his level of intelligence is worthy of special study. Others contend that a practical difficulty with the procedure is that tests of achievement and so-called tests of intelligence really largely measure the same thing, and suggest instead that diagnostic study be given to those pupils whose achievement in some school subject, or subjects, is well below their general achievement level. Still other writers rely heavily upon the judgment of the

¹⁰ See Guy M. Wilson, "Corrective Load in the Fundamentals of Arithmetic in Grades VI, VII, and VIII," pages 234-241 in *The Role of Research in Educational Progress*. Washington, D. C.: American Educational Research Association, May, 1937.

teachers. Baker,¹¹ for example, selected his sixty pupils for special remedial coaching by taking those who had received final marks of failure or conditional passing in four fundamental subjects. He admits that this criterion was used at the outset primarily because of its availability, but states that it "arose steadily in our esteem."

All these suggestions have merit. The judgment of the present teacher should always be taken into account especially since in the ordinary school whatever diagnostic and remedial work is attempted will be undertaken by the regular classroom teacher. But the present teacher's judgment needs to be supplemented by considering the judgment of past teachers as reflected in the school record. Since the judgment of teachers is not infallible however, general achievement tests and intelligence tests will be found particularly valuable. Any pupils in the intermediate grades whose achievement falls a year or more below their age or grade level should usually merit some study. Discrepancies between achievement and intelligence are of particular significance when intelligence has been measured by individual tests or performance tests rather than by ordinary group tests. Such discrepancies also assume added significance when the pupil has apparently had ample opportunity for learning.

While special study and treatment are often justified for the lowest 5 or 10 per cent in the typical class, it must not be thought that diagnosis should be restricted to low ranking pupils and to obvious misfits. On the contrary, some of the most profitable cases are those whose achievement is average or even above but is nevertheless well below what appears possible. As a matter of fact Hildreth¹² points out that many clinics prefer not to attempt remedial work with very dull pupils, say those with IQ's of approximately 80 and below, but prefer instead to alter the achievement goals for such children. It will be found at times that pupils whose personality defects interfere with satisfactory social adjustment have superior academic achievement. In fact psychiatrists point out that the teacher should often be most concerned about the mental health of those who give her least concern academically. The writer recalls the case of a sixth grade girl whose scholastic achievement was well above the norms on the tests but whose attempts at social adjustment to the group had been distinctly unsuccessful. The girl told her mother that she would give anything in the world if she had just one friend. In the conventional school this girl would have been regarded as making an entirely satisfactory record, but in the modern school she is seen to be so seriously maladjusted as to require special attention.

Locating the nature of the difficulty. After locating the pupils who are experiencing trouble, the next step is to make a careful examination

¹¹ Harry F. Baker, *op. cit.* pages 9-16.

¹² Carolyn Hildreth, *Locating the Three R's* (Second Edition), p. 16. Philadelphia: Educational Publishers, Inc., 1947.

of the difficulty of each pupil. A bill of particulars is needed. It is just here that diagnostic tests, if available, are of great value. The aim of such tests is to reveal the specific location of the pupil's difficulties. As a rule, each test has a limited scope, but attempts to explore thoroughly this restricted area. For example, one test might undertake to find the particular number combinations which are causing trouble in the addition of whole numbers, while another test attempts to find out whether inadequate reading ability, faulty technique of analysis and procedure, lack of skill in the fundamental processes, or some other factor is responsible for poor performance in reasoning problems.

Most of the diagnostic tests published to date are limited to the tool subjects mainly on the elementary level. Traxler¹² prepared a comprehensive bibliography of available tests together with a practical discussion of their effective use. Blair¹⁴ compiled similar information with special reference to the high school. Traxler¹³ offered this warning: "Our experience at the Educational Records Bureau indicates that, at present, there is scarcely one test which gives us as much reliable information as is needed for effective diagnosis in any one field."

But any test, whether standardized or not, can be used to reveal the location of errors. The principal advantages of the standardized test are that in content it is likely to represent a more careful selection than the informal test, and that the existence of comparable forms makes it possible to verify the accuracy of diagnosis based on one form and to check upon the success of any remedial measures undertaken. However, these special values in standardized tests by no means rule out the values of informal tests when used for diagnostic purposes.¹⁶ In reading, for example, some writers regard informal tests as even more important than standardized tests. The diagnostic value to be realized depends more upon the teacher than upon the test used. Durrell estimates that at least 75 per cent of the cases requiring special attention in reading can be handled adequately by well trained classroom teachers using non-standardized tests supplemented by observation of the pupils' achievement and work habits. He says:

Such informal tests and observation charts usually indicate the correct level on which to start remedial instruction, the specific reading abilities in which the child is weak, and the faulty habits and confusions which must be overcome in the remedial program.¹⁷

Figure 41 illustrates a useful procedure for analyzing the errors revealed

¹² Arthur E. Traxler, *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects*, 80 pages, New York: Educational Records Bureau, 1942.

¹⁴ Glenn Myers Blair, *op cit*.

¹³ Arthur E. Traxler, *Individual Evaluation*, in *New Directions for Measurement and Guidance*, page 28, Washington: American Council on Education, 1944.

¹⁶ Donald D. Durrell, *Improvement of Basic Reading Abilities*, page 18, Yonkers: World Book Company, 1940.

¹⁷ *Ibid*, page 296. Quoted by special permission.

NAME OF PUPIL	WHOLE NUMBERS										FRACTIONS																													
	Addition					Subtraction					Multiplication					Division					Addition					Subtraction					Division									
Peggy	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
Mildred																																								
James																																								
Betty W																																								
Billy																																								
Dorothy																																								
Istler																																								
Ruth D																																								
Mary																																								
Betty B																																								
Bobby																																								
Ann																																								
Jeanne																																								
Nancy																																								
Marybell																																								
Ruth L																																								
Howard																																								
John																																								
Ben																																								
Lewis																																								
Eraser																																								
Ewing																																								
Julia																																								
Nelle																																								
Maile																																								
Dick																																								
Sally																																								
Sam																																								
SUMMARY																																								
No Errors																																								
No Omissions																																								

Figure 11 Analysis Sheet of Test 3 Metropolitan Achievement Tests Form 1 Arithmetic Fundamentals for a Fifth Grade Class in October (Last 21 Problems Are Omitted)

by a standard test in arithmetic. The procedure is equally applicable to informal tests. This particular test, Test 3, Arithmetic Fundamentals of the Metropolitan Achievement Tests, was administered to a fifth grade in October. The pupils are arranged in descending order according to the score on this test. Each error is indicated by X and each omission by 0 as far as the pupil attempted problems, the problems beyond the last one attempted are indicated by - - - - . The summary at the bottom shows how many times each problem in the test was missed and omitted. This simple analysis reveals clearly what type of problems caused trouble and to whom the trouble was caused. The procedure is really group diagnosis but it may be regarded as the first step in individual diagnosis. It should be apparent that classroom teachers who are content merely to obtain the total score made by each pupil on a test are really overlooking the greatest value of the test for instructional purposes.

Similar error analyses can be made for most subjects, but are especially valuable in mathematics, spelling, reading, handwriting and language. It is usually better to make more than one such analysis, however, than to rely upon a single sampling which is almost sure to include some errors that are merely chance occurrences rather than habitual. Brueckner and Elwell¹⁵ for example, found from the study of a test in the multiplication of fractions containing in random order four examples of each type, that failure to work a single example correctly is hardly a safe index and that at least three problems of each type are required for a valid individual diagnosis. A later study¹⁶ in subtraction showed that all the problems of a type should be grouped together on the test.

It is not sufficient, however, to stop with tabulating the frequencies of questions missed on tests or mistakes made in written work. A further analysis must be made of the types of errors represented. It will be noted that problem 24 in Figure 41 was missed by 23 pupils out of 28. As a basis for remedial instruction the teacher needs to know what types of incorrect solutions were made by her pupils. An examination of the test papers provides the answer. Problem 24 follows:

24 Add

$$\begin{array}{r} 3\frac{1}{3} \\ + 4\frac{1}{6} \\ \hline \end{array}$$



It is found that 15 of the 23 incorrect solutions were $7\frac{2}{3}$; merely a failure to reduce the fraction to its lowest terms. Five of the 6 errors made by the

¹⁵ Leo J. Brueckner and Mary Elwell. Reliability of Diagnosis of Error in Multiplication of Fractions. *Journal of Educational Research* 26: 175-185, November 1932.

¹⁶ Leo J. Brueckner and Mabel J. Hawkinson. The Optimum Order of Arrangement of Items in a Diagnostic Test. *Elementary School Journal* 34: 351-357, January 1934.

best 7 pupils were of this type. Five pupils got as an answer $7\frac{7}{8}$, which represents two types of errors. Still more serious is the status of the pupil who got $\frac{7}{8}$ for an answer. An interesting type of incorrect solution is represented by a pupil whose answer was $7\frac{7}{8}$. It is apparent that he merely added the numerators and the denominators without taking the trouble to reduce the fractions to a common denominator. The other wrong answer was $7\frac{7}{2}$.

A second illustration of the value of error analysis is taken from spelling. A few years ago the writer gave a spelling test to a class of high school seniors. The results were disappointing. One of the words missed most often was "undoubtedly." Contrary to expectation, a tabulation of the errors revealed the fact that the first two syllables were spelled correctly by all pupils. The misspellings were of four forms: "undoubtelly," "undoubtely," "undoubtalv," and "undoubtally." It can be seen that the fundamental error is mispronunciation. The pupils were attempting to *spell* this common word as they were accustomed to *pronounce* it. Hildreth²⁰ reports that confusion over vowels in the middle and end syllables is a prolific source of error, and that syllables containing *e*, *a* and *o* are especially liable to vague, indistinct pronunciation. Another investigator²¹ found that emphasis upon correct pronunciation in reading resulted in a decided improvement in the *spelling of pupils in the fifth and sixth grades*.

One of the greatest values of such error analyses is that they reveal that a relatively few types of errors made over and over again are responsible for the poor performance of most pupils. In an early study of errors in spoken language Charters²² found that 71 per cent of the errors made by Pittsburgh children fell into only five classes. A study in Madison, Wisconsin,²³ revealed that more than half of the total number of language errors made from the kindergarten through the sixth grade represented but four types. In an extensive study Newland²⁴ found that errors in writing only four letters, *a*, *e*, *r*, and *t* accounted for almost half of the illegibilities made, whether by elementary school, high school, or adult groups; and that only four types of difficulties in letter formation caused more than half of the illegibilities. It cannot fail to be encouraging to teachers and pupils alike to find that remedial efforts directed at a relatively few troublesome points may result in great improvement.

Locating the causes of errors. Even more important, and usually far more difficult, than knowing *where* the errors occur is knowing *why* they occur. One limitation of test scores in diagnosis is that they reveal the

²⁰ Gertrude Hildreth *op cit* page 492

²¹ Marjorie E. Kay "The Effect of Errors in Pronunciation upon Spelling" *Elementary English Review* 7: 64-66 March 1930

²² Unpublished report made in 1919

²³ *Language Curriculum Committee Reports* Madison Wisconsin Madison Public Schools 1932

²⁴ T. Ernest Newland "An Analytical Study of the Development of Illegibilities in Handwriting from the Lower Grades to Adulthood" *Journal of Educational Research* 26: 249-258 December 1932

products of learning rather than the learning *process* itself Tyler²³ makes a useful distinction between measurement or appraisal, and interpretation or inference. In other words, causation is not established directly by the act of measurement, but must be inferred from the measurement and other pertinent data. Scates²⁴ puts the situation clearly:

A multitude of test scores are in themselves meaningless. They show facts, but they do not show reasons. They neither diagnose nor evaluate. They may be useful aids, but they leave the principal problem to the teacher's insight, namely, that of determining what is indicated.

At times, as in some of the examples cited, a reasonably safe inference can be made from the nature of the errors themselves. But rarely can a sufficiently complete explanation be made without considering the child's past history, outside the school as well as inside. It is never safe to infer that a child's poor performance in school is due to mental deficiency or personality defects unless a careful study of his educational opportunities has been made. Fortunate indeed is the school whose records are sufficiently complete to provide the essential data.

Certain outstanding physicians and surgeons have advocated an enlarged concept of diagnosis in modern medicine. Several years ago Sir William Osler argued that it was more important to know what kind of man had a certain disease than to know what disease the man had. Wilbur has made the following statement:²⁵

It is just as important in these days for a young doctor to understand his patient's personal life, home responsibilities, and community relationships, as it is to be able to tell just what organisms are living in his lungs or invading his liver. The doctor who has not studied psychology and who cannot acquire a knowledge of it if he is to be successful will have to confine himself to work in the laboratory or be a pure technician.

Hildreth suggests that the following five "areas of investigation"²⁶ are important in diagnosis:

Mental equipment of the learner: Aptitude for academic schoolwork, learning capacity, readiness for learning, habitual modes of response, judgment, reasoning ability, insight, memory, association, perception, attention span, ability to see relationships, creative ability, intellectual interest, suggestibility, comprehension, auto-criticism, habits.

²³ Ralph W. Tyler, *Elements of Diagnosis*, *Thirty-Fourth Yearbook of the National Society for the Study of Education*, page 113. Quoted by permission of the Society, Bloomington, Illinois: Public School Publishing Company, 1935.

²⁴ Douglas E. Scates, "Differences Between Measurement Criteria of Pure Science and of Classroom Teachers," *Journal of Educational Research*, 37:1-13, September 1943.

²⁵ Ray Lyman Wilbur, "The March of Medicine," *Science*, 87:201-202, March 4, 1938.

²⁶ Gertrude Hildreth, *Learning the Three R's*, pages 547-549, Minneapolis: Educational Publishers, Inc., 1936.

Language equipment Command of mother tongue, knowledge of foreign languages language first learned, speech defect, immaturity in speech, in articulation, or diction, vocabulary, rapidity or slowness of speech, history of speech development, age of using words and sentences, descriptive powers, written composition

Personality, temperament, and dynamic equipment Self-control, affability, desirable and undesirable inhibitions, attitudes, friendliness susceptibility, docility, irascibility, drive, perseverance stability, lability of mood compliance, responsiveness restlessness, shyness, tendency toward embarrassment day dreaming fears withdrawal from reality, sex interest, morbid curiosity irrational attitude manners, attitude toward failure and toward the school disability compensations child's interests attitude toward school preferred school subjects child's play interests, obsessions fears worries, ability to get along with other children social qualities attitude toward brothers and sisters and other members of the family delinquent and anti-social activities degree of normal adjustment, changes, growth and development in all these factors since birth

Physical status, sensory and motor equipment physical conditions Sensory acuity, constitutional defects, physical maturation physical handicaps and defects disease history, glandular balance, condition of teeth etiology of illness posture accidents or unusual physical shocks nutrition diet hygiene, psycho-motor status muscular strength or weakness, handedness, steadiness, coordination, efforts to change handedness facility in sports and games

Environment and home history Economic factors, literacy of parents, number of sibs, marital status of parents, foreign background, other adults in the home and their contact with this child, evidences of culture, e.g., books, musical instruments labor-saving devices in the home, harmony in home adjustments, attitude of home toward school, cooperation of home with school, neighborhood environment association with other children, child's opportunity for free time child's activities in free time

Child's daily schedule Rising, eating, sleeping, play, schoolwork at home, regularity or irregularity in home program

School situation, history and present status Methods of instruction especially the work with which the child has difficulty size of class groups, capability of class groups, school marks textbooks and other materials used, progress of other children progress in learning from grade to grade retardation failure or double promotion, attitude of child toward teacher, teacher's usual success with pupils of her grade level, teacher's experience, rapidity with which average child progresses, requirements of the course of study, classification system, provision for individual assistance, date of first recognition of the child's disability, former diagnostic and remedial work carried on with the child both in school and in clinics, survey of all school records that would throw light on the situation, kind and extent of supervision objective test records, analysis of previous training and methods of attack in learning, e.g., to write, evidence of readiness for instruction before work in skills, teacher's story of the case, attitude toward the child discipline in the classroom absence from school, tendency toward tardiness, truancy Information the teacher has about modern methods in education and child study progressiveness of the school program, extent to which teacher makes individual studies and keeps cumulative records of pupil, age of the child on entering school, terms retarded failure in specific subjects, absence Extent to which each teacher is acquainted with the child's past school history, extent to which the teacher knows the facts at the beginning of the school term, teacher's explanation of the cause of difficulty, teacher's recommendations as to what should be done efforts

the teacher has been making to eradicate the difficulty extent to which the teacher capitalizes the child's interests

It is, of course manifestly impossible, as well as usually unnecessary, to consider all these facts in any particular case. Satisfactory explanation of the less serious cases can often be found in a relatively few factors although rarely, if ever, in just one. The more serious cases will usually be found more complex to analyze as well as more difficult to remedy.

It will frequently be necessary to supplement the data of the existing school records. A visit to the pupil's home is often helpful. A careful observation of the pupil at work is another fruitful source of information. Objective records of observations made under controlled conditions are particularly important. Considerable light is often thrown upon the attitudes and work habits of unsuccessful pupils by observing them at work and then by comparing successful pupils under similar conditions.

A skillful interview by a tactful teacher will sometimes give a clue to the difficulty when other methods fail. In the upper grades and the high school, check lists, questionnaires, and other forms of written responses are valuable aids to the personal interview. Having the pupil "think out loud" through the solution of a problem in mathematics or science or give an explanation of the procedure used is often most illuminating.

Two illustrations make clear the value of the interview as a supplement to the written test in locating the sources of difficulty in arithmetic. Buswell tells of a boy of better than average intelligence whose work in column addition was both slow and inaccurate. To the interviewer he explained that he did not like to add and so wanted to get the worst of it over as soon as possible. For this reason he always added the numbers according to size, beginning first with the largest numbers and leaving the smallest ones till last. But as this technique meant skipping up and down the column, it involved great risk of omitting some of the numbers altogether and of adding others more than once. The story is told of a sixth grade school girl who had an elaborate but somewhat ineffective "system" for solving reasoning problems. Her explanation was somewhat like this: "Whenever there's lots of numbers I add but when there's only two numbers with lots of parts [digits] I subtract. But if there is just two numbers and one is littler than the other I divide when they come out even, and multiply when they don't." It is most unlikely that any analysis of test papers or observation of the pupils at work, would have resulted in a correct inference as to the real trouble in either of the cases above.

Teachers often find that an interview with the pupil sheds needed light upon difficulties in reading and English. Pressey and Campbell²² report that one ninth grade pupil explained capitalizing the word "Pirates" on the ground that pirates are real persons just as much as "John Silver" or

²² Sidney L. Pressey and Pera Campbell, "The Causes of Children's Errors in Capitalization: A Psychological Analysis," *English Journal* 22: 197-201, March 1933.

"Captain Kidd" Another teacher discovered that a boy had written "a quarter to three" in answer to a question on a reading test when the correct answer was "twenty five minutes till three" because everybody knows that twenty-five cents make a quarter!

Brownell³⁰ has shown the possibilities of classifying the mental processes used by the pupils as revealed by interviews according to levels of maturity represented. He concludes that a reasonably flexible interview technique in analyzing learning is "exceedingly valuable if it is sagaciously employed." One survey³¹ of the experimental literature relating to the reliability of the interview arrives at the conclusion that "with well trained interviewers working under carefully defined conditions quantitative interview ratings representing a complex over all evaluation can be made as reliable as most personality tests and more reliable than some of them. Nevertheless good interviewing requires skill as well as time and patience."

Remedial procedures The ultimate purpose of diagnosis is to afford a basis for effective remedial procedures. When the cause or causes of the pupil's unsatisfactory adjustments have been determined an intelligent program of correction can be planned and not until then. Whenever the same causes appear to operate in several pupils group measures may be satisfactory. Usually, however, remedial programs must be planned for each pupil individually.

A study by Davis³² shows the close relationship between educational diagnosis and remedial instruction. Two extra periods a week were devoted to 275 pupils of poor spelling ability in grades 2B to 6A inclusive. The results showed "marked improvement." Pupils remained in the remedial classes until they made perfect scores on the spelling tests of two successive Fridays. The average time required was 7.5 hours and bore little relationship either to intelligence or grade location. Twenty-four different types of difficulties were located, and listed with each difficulty were the most successful remedies found by the teachers. The ten most common difficulties with their remedies, are shown in Table 34.

Traxler³³ has prepared some very convenient charts which outline appropriate diagnostic and remedial procedures for common types of disabilities in reading, arithmetic, language usage, spelling and handwriting. Figure 42 shows the chart for handwriting. Note that a detailed analysis of samples of the pupil's writing is suggested as well as diagnostic charts and tests.

³⁰ William A. Brownell, Rate, Accuracy and Process in Learning. *Journal of Educational Psychology* 35: 321-327, September 1944.

³¹ Sidney H. Newman, Joseph M. Bobbitt and Dale C. Cameron, The Reliability of the Interview Method in an Officer Candidate Evaluation Program. *American Psychologist* 1: 103-109, April 1946.

³² Georgia Davis, Remedial Work in Spelling. *Elementary School Journal* 27: 613-626, April 1927.

³³ As listed by Traxler *op cit* pages 34-35.

TABLE 31

DISTRIBUTION OF SPELLING DIFFICULTIES AND SUCCESSFUL REMEDIES (AFTER DAVIS)

Difficulties and Remedies	Frequency
1 Has not mastered the steps in learning to spell a word a Teach steps until every child knows them and uses them b Study each word with the children	88
2 Writes poorly a Discover particular letters or combinations of letters that are difficult and practice on these letter combinations b Practice words containing writing difficulties	88
3 Cannot pronounce the words being studied a Go over the words before the children study them so that every child will know what he is studying b Help the child to unlock words for himself	78
4 Has bad attitude toward spelling a Supervise study closely so that the child will get into the habit of studying words correctly without wasting time b Try to show need for study c Give study work under time pressure d Try to appeal to pride e Try to work up competition with self (that is, of the pupil with himself) f Give reward	71
5 Does not associate the sound of the letters or the syllables with the spelling of the word a Teach letter sounds b Listen to careful pronunciation c Teach the child to syllabify words d Say words slowly again and again to hear sounds	49
6 Needs more time than can be devoted to spelling in the regular class a Give more time after school or during the day when other work is finished	21
7 Is discouraged because he misspelled so many words in the Monday test a Take a few words at a time b Study at odd times during the day c Have the pupil stay longer in the afternoon than the others	20
8 Has speech defect a Listen to pronunciation b Look at word carefully c Teach difficult combinations	16
9 Does not mark paper correctly a Teach child how to check b Insist on rechecking c Always check paper	16
10 Interchanges letters a Study words carefully b Underline difficult part c Try to spell by syllables	10

An effective method with bright pupils may fail with dull. In fact, no method is likely to improve materially the academic achievement of the mentally deficient child. Even with normal or superior children the substitution of correct habits for incorrect will require time. No sudden trans-

CHART V HANDWRITING
SUGGESTED DIAGNOSTIC AND REMEDIAL PROCEDURES

TYPE OF DEFECT	DIAGNOSTIC PROCEDURE	SUGGESTED TYPES OF REMEDIAL TREATMENT
1 Slant a Too much slant b Writing too straight c Lack of uniformity	1 Use diagnostic chart study different samples of writing. Draw lines through letters parallel to slant on different parts of page. Compare these lines as to direction. Observe pupil as he writes and note details--position paper etc.	1 Some instances of poor slant can be corrected by changing position of writing arm or manner of grasping pen. Change in position of paper will help others. Note that paper should be at an angle. Other pupils must learn to turn their hand as they approach end of line. Explain to pupils effect of slant on quality.
2 Alignment a Lack of uniformity b All letters at not the same height	2 Use diagnostic chart draw horizontal lines through writing even with top and bottom of some of the letters.	2 Explain defect to pupil. Lack of uniformity of alignment results partly from motor inco-ordination and will probably be corrected as co-ordination of writing movements improve through practice.
3 Quality of line a Writing too heavy b Writing too light c Line wavy and uncertain	3 Use diagnostic chart note type and size of pen and manner of holding it note speed of writing.	3 Make sure that pupil has proper writing materials see that he does not use his writing arm to support his body. If line is thin and wavering give drills to speed up movement and improve co-ordination.
4 Formation of letters a Poor general form b Lack of smoothness c Parts omitted d Parts added e Letters not closed	4 Use diagnostic chart if desired letter form may be analyzed in detail with Pressey chart. Study general form and habits of forming each letter. Often faults in letter form are related to only a few letters.	4 Make some use of movement drills to improve smoothness. Practice especially on movements common to several letters. Study details of letter form with pupils and show them where they need to improve. Have pupils practice individually on the letters which diagnosis has shown to be poorly formed.

CHART V (Continued)

TYPE OF DEFECT	DIAGNOSTIC PROCEDURE	SUGGESTED TYPES OF REMEDIAL TREATMENT
5 Spacing of words a Too wide b Too narrow c Not uniform	5 and 6 Use diagnostic chart, study various samples note whether wide spacing or crowding occurs on different part of page Observe pupils while writing for evidence of too much lateral movement	5 and 6 Explain fault to pupil Have him pay special attention to spacing while writing samples to be inspected by teacher Movement exercises are of some value in improving spacing
6 Spacing of letters a Too wide b Too narrow c Not uniform		
7 Size of writing a Too large b Too small c Lack of uniformity	7 Study different samples and compare with those of other pupils in the same grade Considerable variability is allowable among individuals and especially between grades Young pupils tend to write large Note freedom of movement Try to discover cases of lack of uniformity	7 Writing that is too small may result from a cramped finger movement Give movement exercises to relax pupil and bring about some arm movement If writing is too large pupil can sometimes correct it through conscious effort if his attention is called to it In young pupils improvement may have to await the process of maturation
8 Writing not neat a Blotches b Words crossed out and rewritten	8 Examine samples of writing especially those prepared in daily work With respect to blotches see if writing materials are defective	8 See that pupil has proper writing materials and that they are kept in working order Explain effect of lack of neatness on all school work Make daily work in other subjects the gauge of neatness
9 Speed a Writing too slow b Writing too fast	9 Speed of writing affects the quality, but aside from this fact it is important in that some pupils write so slowly and laboriously that they have difficulty in preparing assignments on time Give a test of speed of writing and compare number of letters per minute with grade norms	9 If writing is too fast show pupil its effect on letter form and have him write samples under timed conditions Some pupils go to the other extreme and write so slowly that they practically draw the letters Give movement exercises while counting rapidly to break down habits of slow movement Insist that pupils speed up writing regardless of their letter forms Their writing will probably deteriorate for a time, but when old habits are broken down teacher and pupil can build new ones

Figure 42 Traxler Chart of Suggested Diagnostic and Remedial Procedures in Handwriting

formation is to be expected. But if only negligible progress results from extended practice, the remedial program should be revised.

Preventive diagnosis. In the long run, the greatest value of a diagnostic and remedial program is the discovery of preventable factors within the control of the school which lead to maladjustment. Frequently modifications in school organization, curriculum, instructional materials, and teaching methods are suggested by an analysis of what is happening to the pupils under the existing program. Manifestly, factors which have produced learning difficulties in the past are likely to do so in the future. It is always better, and generally easier, to prevent errors than to correct them. It will often be found that a program of studies which provides wider differentiation in method and content to suit pupils of varying abilities and interests is the way out of many difficulties. The systematic use of readiness tests of various types to determine when the pupil is sufficiently mature physically, mentally, and socially to begin the regular work of the first grade and a judicious use of aptitude tests to establish the pupil's fitness for the more formal and abstract subjects such as arithmetic, algebra, and foreign language will prevent much needless failure. Terman³⁴ says: Perhaps the most important conclusion to be drawn from the extensive researches here reported is that disability of any degree in any of the basic school subjects is wholly preventable. 'Prevention is the highest level of diagnosis: its ultimate goal.'

SELECTED REFERENCES FOR FURTHER READING

- Betts, Linnett Albert. *Foundations of Reading Instruction with Emphasis on Differential Guidance*. New York: American Book Company, 1946. 758 pages.
 Blair, Glenn Myers. *Diagnostic and Remedial Teaching in Secondary Schools*. New York: The Macmillan Company, 1946. 422 pages.
 Boyd, Gertrude and Schwiering, O. C. A Survey of Child Guidance and Remedial Reading Practices. *Journal of Educational Research* 43: 494-506, March 1950.
 Cook, Walter W. The Functions of Measurement in the Facilitation of Learning. Chapter 1 in I. F. Lindquist (Editor), *Educational Measurement*. Washington, D. C.: American Council on Education, 1951.
 Cronbach, Lee J. *Educational Psychology*. New York: Harcourt, Brace and Company, 1954. Chapters 6 and 7: Assessing Readiness, I: Personality and Motivation and Assessing Readiness, II: Abilities.
 Dressel, Paul L. and Mann, William A. Appraisal of the Individual. *Review of Educational Research* 21: 115-131, April 1951.
 Hildreth, Gertrude. *Learning the Three R's* (Second Edition). Philadelphia: Educational Publishers, Inc., 1947. 897 pages.
 McCullough, Constance M., Strang, Ruth M., and Traxler, Arthur E. *Problems in the Improvement of Reading*. New York: McGraw-Hill Book Company, 1946. 406 pages.

³⁴ Lewis M. Terman. Foreword to Grace M. Fernald's *Remedial Teaching in Basic School Subjects*, page ix. Copyright 1943. Reprinted by permission of the publishers, McGraw-Hill Book Co.

- Simpson, Robert G , *Fundamentals of Educational Psychology* Philadelphia J B Lippincott Company, 1949 Chapter 13, "Analyzing the Learner's Difficulties "
- Stauffer, Russell G , "Certain Basic Concepts in Remedial Reading," *Elementary School Journal*, 51 331-342, February, 1951
- Stuit, Dewey B "Counseling Methods Diagnostics," *Annual Review of Psychology*, 2 305-316, 1951
- Tyler, Ralph W , "The Functions of Measurement in Improving Instruction," Chapter 2 in F F Lindquist (Editor), *Educational Measurement* Washington, D C American Council on Education 1951

13

Classification and Promotion

A. The Nature and Educational Significance of Human Variability

The problem of human variability. The existence of variability is one of the best established facts about human beings. Obvious differences in height, weight, strength, and good looks could hardly escape the notice of the most casual observer. The greatest seers and wise men of all ages have recognized also the less obvious but more important differences in ability, interests, and needs. One of the familiar parables of Jesus, for example, is that of the talents.¹

It would be difficult today to find a fuller recognition of the educational significance of individual differences than appears in the writings of those two apostles of human liberty, Jean Jacques Rousseau and Thomas Jefferson. Rousseau asserted that "it would be a great mistake to bestow it [instruction] on all children indiscriminately and without regard to their individual differences."² Jefferson wrote of a proposed educational measure "The general objects of this law are to provide an education adapted to the years, capacity and the condition of every one, and directed to his freedom and happiness."³ It is apparent, therefore, that when the author of the Declaration of Independence penned the famous line, "All men are created equal," he had in mind *equality before the law*, and that he recognized fully the duty of the state through education to provide *equality of opportunity*. A prominent American educator⁴ argues that "our concepts

¹ "And unto one he gave five talents to another two, and to another one to every man according to his several ability." Matthew 25:15

² Jean Jacques Rousseau *The New Heloise*, Part V, Letter 3

³ Thomas Jefferson *Notes on Virginia* pages 250-252

⁴ I. Newton Edwards, "We Need New Purposes in Education," *Phi Delta Kappan* 28:16 September, 1946

of freedom and equality are outmoded" and cannot both be realized, since they "are in fact mortal enemies."

It is surprising, therefore, to find that the problem of individual differences was not seriously treated in psychology before the time of Galton in the latter half of the nineteenth century, a neglect which has been characterized as perhaps the "most extraordinary blind-spot in previous psychology."⁵

Otto⁶ estimates that during the last twenty years more time and effort in educational research have been devoted to the study of individual differences than to any other single topic; he is greatly impressed with the extensive literature available. Yet in 1925 a competent school psychologist stated that the "schools heretofore have to a large extent ignored these differences."⁷ Five years later a national survey revealed that "provisions for individual differences, in general, are innovations in the secondary schools."⁸ In 1936 a survey of 300 courses of study showed that only about one in ten "contain any suggestions for adapting instruction to individuals."⁹ Eight years later an educational psychologist¹⁰ characterized as largely "lip service" the attention educators give to individual differences. Davis says:

Despite its philosophy of individualization, the school, in practice, fosters a program of regimentation and standardization.

In the meantime, better enforcement of the compulsory education laws and the rapid increase of secondary-school enrollments have served but to intensify the problem, the nature of which was more accurately revealed by scientific measurement.¹¹

Group differences. Scientific research, on the whole, has shown that differences between groups are not so great as they are commonly assumed

⁵ Gardner Murphy, *Historical Introduction to Modern Psychology* (Revised Edition), page 117 New York: Harcourt, Brice & Company, Inc., 1949.

⁶ Henry J. Otto, *Elementary School Organization and Administration* (Second Edition), page 160 New York: D. Appleton Century Company, 1944.

⁷ A. A. Sutherland, "Factors Causing Maladjustment of Schools to Individuals," *Twenty Fourth Yearbook of the National Society for the Study of Education, Part II*, pages 29-30 Bloomington, Illinois: Public School Publishing Company, 1925.

⁸ Roy O. Billett, *Provisions for Individual Differences, Marking and Promotion*, National Survey of Secondary Education, Monograph No. 13, page 8 Washington, D. C.: United States Office of Education, 1932.

⁹ Henry Harap, "Differentiation of Curriculum Practices and Instruction in Elementary Schools," *Thirty Fifth Yearbook of the National Society for the Study of Education, Part I*, page 162 Bloomington, Illinois: Public School Publishing Company, 1936.

¹⁰ Robert A. Davis, "Experimenting in Education," *Educational Administration and Supervision*, 30: 1-16 January, 1944.

¹¹ For an excellent summary, see A. R. Gilliland and E. L. Clark, *Psychology of Individual Differences*, 535 pages New York: Prentice-Hall, Inc., 1939. A standard work in this area is Anne Anastasi and John P. Foley, Jr., *Differential Psychology: Individual and Group Differences in Behavior* (Revised Edition), 894 pages New York: The Macmillan Company, 1949.

to be. There is little basis for the widespread illusion that the group of which one happens to be a member is superior, while all others are inferior. The intellectual differences between the sexes, for example, are in general slight. Furthermore, all levels of mental ability are found in all economic, occupational, and social groups, although not in the same proportions. Even the differences between races have been grossly exaggerated, and such differences as appear probably reflect cultural rather than innate intellectual variations. It is manifestly impossible to make adequate provision for individual differences by classifying pupils for instructional purposes according to the social, economic, occupational, racial, or other similar group from which they come. Fortunately, perhaps, for democracy the problem is not so simple as that.

Almost without exception the average differences between groups are less significant than the differences within any single group. An important example of this is the enormous overlapping among school grades. Although the average difference in intelligence and in achievement between successive school grades rarely exceeds one year, the difference within any grade is likely to be at least four or five years. As a matter of fact, Baker¹² points

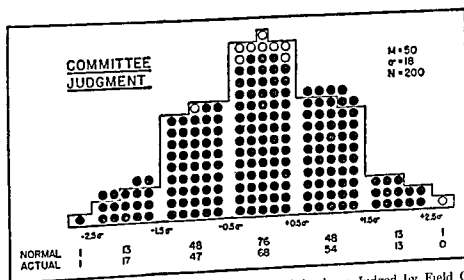


Figure 13. General Quality of 200 Secondary Schools as Judged by Field Committees. Each ● represents an actual school. Each ○ represents a school in theoretical distribution. (From *Education of Secondary Schools, General Report*, page 110.)

out that the achievement of the more capable halves, or the less capable halves, of two adjacent grades is usually much more alike than that of the two halves of the same grade. On a test of general academic knowledge, the Pennsylvania study showed that about 10 per cent of the high school seniors exceeded the median of the college seniors, while nearly 10 per cent

¹² *Thirty-Fifth Yearbook*, *op. cit.*, pages 137, 145.

of the college seniors fell below the median of the high-school seniors.¹³

It must not be thought, however, that one group is just like every other. As a matter of fact, certain types of groups differ from each other very much as the individuals within any one group differ from each other. For example, Figure 43 shows the distribution of the ratings of 200 high schools, which closely approximates the normal curve. It is quite likely that all the schools in a single state would be similarly distributed on practically every characteristic.

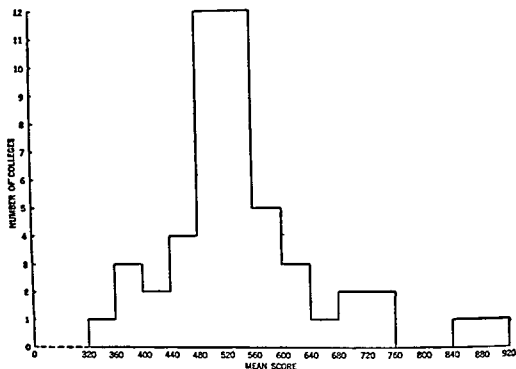


Figure 44. Distribution of Mean Scores of Seniors in Forty-Nine Colleges in Pennsylvania on a Test of General Academic Knowledge. (Data from *The Student and His Knowledge*, page 78.)

Figure 44, although somewhat asymmetrical, shows that, on the basis of the mean achievement of their seniors, 49 colleges in Pennsylvania have the wide range and the heavy concentration near the center that characterize normal curves. In other words, there are differences among institutions just as there are among individuals. It is this fact that makes the traditional classification of schools for accrediting purposes such a baffling problem, and that has been responsible for the trend toward evaluating each school in relation to its own objectives and program rather than in relation to other schools. It is now being recognized that it is just these differences that give individuality and distinction to institutions.

¹³ William S. Learned and Ben D. Wood, *The Student and His Knowledge*, page 21. New York: Carnegie Foundation for the Advancement of Teaching, 1938. For a later illustration from the results of the Army General Classification Test (AGCT), see: Walter V. Bingham, "Inequalities in Adult Capacity—from Military Data," *Science*, 104: 147-152, August 16, 1946.

Individual differences. In contrast with the differences between groups, which have frequently been overestimated, the differences within the group have usually been underestimated. While a vague notion of individual differences has long been in existence, no adequate knowledge of the nature and extent of these differences was possible before the appearance of scientific measurement. Such profound thinkers as Plato, for example, believed that all persons fell into a few rather distinct groups. In fact, the idea that many human abilities are distributed on a continuum, with a concentration near the middle, is a modern conception.

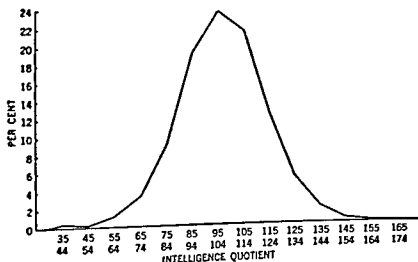


Figure 15. Distributions of Composite IQ's on Forms L and M of the Revised Stanford Binet Intelligence Scales for a Standardization Group of 2,904 Individuals of CA's 2 to 18 years (From Terman and Merrill *Measuring Intelligence* page 37)

Figure 45, according to Terman and Merrill, "probably gives the clearest picture available of the intellectual differences which obtain among American-born white children of the ages in question"¹⁴ Figure 21 on page 260 shows a similar distribution for ninth-grade pupils. Three characteristics of the so called "normal curve" should be noted (1) the *wide range* from lowest to highest scores, (2) the *continuous distribution*—no breaks, and (3) the *distinct tendency to pile up near the center*. Many distributions of test scores have these same characteristics. Skewed curves differ from symmetrical curves in that the heavy concentration is not exactly at the center. In a normal distribution approximately two thirds of the pupils lie within a standard deviation distance from the mean. After a survey of the experimental evidence, Hull says¹⁵

¹⁴ Lewis M. Terman and Maud A. Merrill *Measuring Intelligence* page 37 Boston Houghton Mifflin Company, 1937

¹⁵ Clark L. Hull *Aptitude Testing* page 36 Yonkers New York World Book Company, 1928

We shall probably not be in great error if we conclude that *among individuals ordinarily regarded as normal, in the average vocation the most gifted will be between three and four times as capable as the poorest*

Important as is the wide range of ability between the two extremes, the importance to education of the *continuous distribution* is equally great. On no trait do individuals naturally fall into a few distinct groups, such as 'inferior,' 'average' and 'superior,' or 'dull' 'normal,' and 'bright'. Such so-called "types" are purely arbitrary. It would be possible to make an equally good case for any other number of classes. "In a literal sense everyone is exceptional."¹⁶ There are similar differences in nonintellectual traits.

Trait variability. Not only are there differences among groups, and differences among the individuals of any one group, but there are also important differences among the traits making up any particular individual. Hull made a careful study of these differences and came to the conclusion that "the distribution of talent within an individual follows the normal law much as do the distributions of individual differences."¹⁷ Not only did he observe a "distinct tendency to approach the characteristic shape of the normal probability curve," but he also found evidence that 'the average individual's best vocational potentiality must be between two and one-half and three times as good as his worst'.¹⁸ The importance of these trait differences in an individual for educational and vocational guidance can hardly be overemphasized. It is also apparent that satisfactory ability grouping in one trait may be wholly unsatisfactory in other traits.

The educational problem is further complicated by the fact that, while intercorrelations of these traits are usually positive, the correlations are far from perfect. This means that when an attempt is made to secure a group homogeneous in one factor, it is still heterogeneous with respect to other factors. It is apparent, therefore, that it is impossible to make groups truly homogeneous for instructional purposes, even if it were desirable to do so. The best that can be done is to reduce the amount of heterogeneity. Opponents of ability grouping have made much of this point, apparently quite oblivious to the fact that *ipso facto* they are attacking a straw man, for, manifestly, one need shed no tears over the dangers of an educational situation which one's own data prove to be a physical impossibility.

The concept of the versatile individual who is equally gifted in a considerable number of directions is largely a fiction and as an educational ideal is capable of doing much harm. It has been ridiculed as follows:¹⁹

¹⁶ Edmund S. Conklin and Frank S. Freeman *Introductory Psychology for Students of Education* page 515 New York: Henry Holt & Company, 1939.

¹⁷ Clark L. Hull *op cit*, page 46.

¹⁸ *Ibid* pages 46-49.

¹⁹ Amos E. Dolbear 'Antediluvian Education' *Journal of Education* 68:424 1908.

In interglacial times while the animal kingdom was being differentiated into swimmers, climbers, runners, and fliers, there was a school for the development of the animals.

The theory of the school was that the best animals should be able to do one thing as well as another.

If an animal had short legs and good wings, attention should be devoted to running, so as to even up the qualities as far as possible.

So the duck was kept waddling instead of swimming. The pelican was kept waddling in short wings in the attempt to fly. The eagle was made to run and allowed to fly only for recreation.

All this in the name of education. Nature was not to be trusted for individual should be symmetrically developed and similar for their own welfare as well as for the welfare of the community.

The animals that would not submit to such training but persisted in developing the best gifts they had, were dishonored and humiliated in many ways. They were stigmatized as narrow minded and specialists, and special difficulties were placed in their way when they attempted to ignore the theory of education recognized in the school.

No one was allowed to graduate from the school unless he could climb, swim, run, and fly at certain prescribed rates, so it happened that the time wasted by the duck in the attempt to run had so hindered him from swimming that his swimming muscles had atrophied, and so he was hardly able to swim at all, and in addition he had been scolded, punished, and ill treated in many ways so as to make his life a burden. He left school humiliated, and the *ornithorhynchus* could beat him both running and swimming. Indeed, the latter was awarded a prize in two departments.

The eagle could make no headway in climbing to the top of a tree, and although he showed he could get there just the same, the performance was counted a demerit, since it had not been done in the prescribed way.

An abnormal eel with large pectoral fins proved he could run, swim, climb trees, and fly a little. He was made valedictorian.

Educational provisions for individual differences. Attention has already been called to the fact that few schools are making adequate provisions for the individual differences existing in their pupils. No point stood out more prominently in Billett's study²⁰ than this. Table 30 summarizes the situation for secondary schools in 1930. Billett reduces these provisions to seven categories: (1) homogeneous grouping, (2) special classes, (3) plans characterized by the unit assignment, (4) scientific study of problem cases, (5) variation in pupil load, (6) out of school projects and studies, and (7) advisory or guidance programs. Of these the first three have been found to be core elements in a typically successful program to provide for individual differences.²¹ But it will be noted from the last column that the most successful provision in the opinion of those using it was homogeneous grouping, which had a ratio of 26 per cent. In other words, hardly more than one principal in four or five using any of these plans has a considerable degree of confidence in them.

²⁰ Roy O. Billett, *op cit.*, pages 8-11.

²¹ *Ibid.*, page 11.

TABLE 35

FREQUENCIES WITH WHICH VARIOUS PROVISIONS FOR INDIVIDUAL DIFFERENCES WERE REPORTED IN USE OR IN USE WITH UNUSUAL SUCCESS, BY 8 594 SECONDARY SCHOOLS IN 1930 (AFTER BILLET)

Nature of Provision	Provision in Use		Provision in Use with Estimated Unusual Success		Ratio of Number of Provisions in Use to Number in Use with Estimated Unusual Success
	Num ber	Per Cent	Num ber	Per Cent	
1 Variations in number of subjects a pupil may carry	6 428	75	795	9	12
2 Special coaching of slow pupils	5 099	59	781	9	15
3 Problem method	4 216	49	444	5	10
4 Differentiated assignments	4 047	17	788	9	20
5 Advisory program for pupil guidance	3 604	42	540	6	15
6 Out-of-school projects or studies	3 451	40	439	5	13
7 Homogeneous or ability grouping	2 740	32	721	8	26
8 Special classes for pupils who have failed	2 612	30	350	4	13
9 Laboratory plan of instruction	2 611	30	323	4	12
10 Long unit assignments	2 312	27	349	4	15
11 Project curriculum	2 293	27	365	4	16
12 Contract plan	2 293	27	465	5	20
13 Individual instruction	2 115	25	309	4	14
14 Vocational guidance through exploratory courses	1,911	22	186	2	10
15 Educational guidance through exploratory courses	1 900	22	193	2	10
16 Scientific study of problem cases	1 343	16	146	2	11
17 Psychological studies	1 077	12	70	1	06
18 Opportunity rooms for slow pupils	946	11	172	2	18
19 Morrison plan	737	9	175	2	24
20 Special coaching to enable capable pupils to skip a grade or half grade	726	8	114	1	16
21 Promotions more frequent than each semester	686	8	103	1	15
22 Remedial classes or rooms	593	7	90	1	15
23 Adjustment classes or rooms	544	6	55	1	10
24 Modified Dalton plan	486	6	52	1	11
25 Opportunity rooms for gifted pupils	322	4	69	1	21
26 Restoration classes	191	2	24	0	13
27 Dalton plan	162	2	15	0	09
28 Winnetka technique	119	1	14	0	12
29 Other techniques	101	1			

Table 36, based on returns from 48 large and 58 small cities, shows trends in the elementary school.²² Increasing attempts to introduce more flexible

²² V. V. Caldwell, 'Some Facts Regarding Elementary School Trends', *School and Society* 49:285-288, March 4, 1939.

TABLE 36

TRENDS TOWARD GREATER PROVISIONS FOR INDIVIDUAL DIFFERENCES IN
ELEMENTARY SCHOOLS (AFTER CALDWELL)

<i>Provision</i>	<i>Large Cities</i>		<i>Small Cities</i>	
	<i>Num ber</i>	<i>Per Cent</i>	<i>Num ber</i>	<i>Per Cent</i>
A Daily Schedule				
1 Longer period	23	50	32	55
2 Flexible program	39	85	49	85
3 Subject matter headings eliminated	17	37	22	38
4 Skills content creative activities	35	76	38	66
B Curriculum Content				
1 Child experiences as learning basis	38	83	43	74
2 Elimination of specific subjects	9	20	15	26
3 More freedom for teacher in interpreting course of study	39	85	50	86
4 Emphasis on habits not fact-learning	34	74	35	60
5 Elimination of drill periods	7	15	13	22
6 Relating learning materials to maturation of child	34	74	37	64
7 Relating learning material to immediate need and mental capacity	39	85	38	66
8 Experience used to develop number concepts	32	70	31	53
9 Delay in formal presentation of abstract arithmetic facts	27	59	37	64
10 Elimination of health as a subject	25	54	33	57
11 Provision for hobby development	39	85	44	76
12 Vacation activity program development	24	52	28	48
C Physical Environment				
1 Comfortable adjustable school furniture	32	70	42	72
2 Automatic lighting equipment	11	24	13	22
3 Automatic heating control	30	65	30	52
4 Materials used for sight conservation	21	45	19	33
5 Provision for lunch room	30	65	29	50
6 Provision for rest facilities	28	61	26	45
7 Isolation of sick children	24	52	29	50
8 More floor space per child	17	37	15	26
9 Provision for safe play apparatus	25	54	33	57
10 Provision for ample playgrounds	37	80	41	71
11 Provision for play space in bad weather	18	39	25	43
D Materials				
1 Basic texts eliminated (skills content)	10	22	16	28
2 Wide reading material various levels	40	87	55	95
3 Elimination of work books etc	12	21	25	43
4 Variety of material for creative work	40	87	50	86
E Classification				
1 Provision for pre-school clinics	32	70	38	66
2 Use of reading readiness tests	38	83	37	64
3 Delay in beginning reading program	29	63	29	50
4 Grouping by social age rather than by intelligence or achievement	17	37	15	26
5 Use of no failure program	12	26	16	28
6 Reduction in pupils retained	35	76	41	71
7 Use of tests for guidance not promotion	37	80	46	79
8 Reduction in number of pupils per teacher	19	41	27	47

educational programs are shown. It is also apparent that much yet remains to be done. But it is encouraging to note that somewhat more than a third of these schools were using reading-readiness tests and other tests for diagnostic and guidance purposes rather than merely as a basis for promotion.

B. The Activity Movement

In recent years no program of instruction has received more attention among educators than the activity movement, usually a prominent feature of the so-called "progressive schools." Yet educational historians assure us that the principle that man learns by doing is "as old as man's earliest education."²² In fact, its roots lie further back than the beginning of formal education in schools. Its advocates go even further and assure us that it is grounded in the fundamental nature of the learner himself. However, there are such wide divergencies among its champions, both in theory and in practice, that it may be said that the activity movement not only recognizes individual differences to an astonishing degree, but also actually demonstrates such differences. The essential features of this educational program may be briefly, and somewhat inadequately, described as follows:

1. Education results from the child's own purposeful activity with processes considered personally vital to him. An *activity*, according to Kilpatrick, is "a unitary sample of actual child living as nearly complete and natural as school conditions will permit."²³ At every stage the organism reacts as a whole and the physical, intellectual, and emotional experiences are interrelated.

2. Learning is inherent within the life process itself. It results naturally from the learner's self-directed purposeful activity. Teaching, like learning, is individual in character, arising from a felt need. The teacher is only a guide, and all subject matter is merely a tool. The activity program clearly places upon the shoulders of the classroom teacher the difficult problem of adjustment to individual differences.

3. Interest is at all times the motivating factor in the learning process. Although all teaching procedures recognize the value of interest, the activity movement emphasizes more than any other program the importance of inner drives and interests of the individual pupil, as opposed to extraneous motivation of any kind.

4. The development of the learner's personality, rather than the accumulation of facts and skills, is the objective of all learning. The personality of each individual will develop in accordance with his own abilities, interests, and personal experiences.

5. The evaluation of this relatively intangible personal development is

²² Thomas Woody, 'Historical Sketch of Activism,' in *Thirty-Third Yearbook of the National Society for the Study of Education, Part II*, pages 9-43. Bloomington, Illinois: Public School Publishing Company, 1934. Quoted by permission of the Society.

²³ *Thirty Third Yearbook, op. cit.*, page 63. Quoted by permission of the Society.

volves a fairly long time-span and, therefore, lends itself more to qualitative than to quantitative judgment. In the evaluation process the pupil himself is an active participant. According to Dewey, "the more mature and experienced the teacher, the less will he or she be dependent upon tangible, directly applicable external tests, and will use them, not as final, but as guides to judgment of the direction in which development is taking place."²³

It should not be overlooked however that regardless of the relative emphasis, such activities as reading and arithmetic are always going to be important, and there appears to be no good reason to rely entirely upon subjective impressions when objective measures are available. The mere fact that adequate measures of the less tangible outcomes are not yet available is no justification for neglecting the measurement of the tangibles the tools for which do exist. Furthermore, the absence of suitable tools no more removes the need for evaluation than a lack of food relieves the pangs of hunger. Indeed, the need is probably greater, as Gates suggests.²⁴

Any scheme of education that emphasizes the nature and needs of the individual child, as most progressive programs do, has far greater need of measurements than conventional programs designed primarily to impart information and skill to pupils *en masse*.

C. Homogeneous or Ability Groups

Individual and group instruction. It has sometimes been erroneously assumed that there is a necessary conflict between individual and group instruction. While all learning is individual learning, it can take place in a group setting, and certain types of learning can take place only in a group setting, for the individual not only learns in the group, he learns *from* the group as well. In other words the important question is: What kind of *group organization* best provides for *individual learning*? The problem is to find somewhere between the two extremes of a complete tutorial system and an out and out lecture system the program which represents the best possible compromise between that which is educationally ideal and that which is administratively feasible.

Homogeneous or ability groups. Shortly after the development of group intelligence tests in 1917, educational leaders began to use these tests for grouping pupils in school. This procedure was commonly referred to as "homogeneous grouping." It soon became evident, however, that such groups were far from homogeneous even in intelligence, not to mention other characteristics. The best result that can be obtained under ordinary school conditions is to reduce somewhat the heterogeneity of the instructional groups. The term "ability grouping" came into use as a more accurate term, although frequently used interchangeably with "homogeneous

²³ *Thirty Third Yearbook of the Society* page 83. Quoted by permission of the Society.
²⁴ *Thirty Third Yearbook of the Society* page 164. Quoted by permission of the Society.

grouping " While much confusion still exists, many writers have recently attempted to make a distinction between these terms Instructional groups which are made less heterogeneous in learning ability, usually by the employment of general intelligence tests, are called "ability groups " Groups formed upon the basis of some common interest, social maturity, or other similar basis, are called "homogeneous groups " An activity in a progressive school, although made up of pupils of varying abilities, is certainly homogeneous from the standpoint of the objective sought Most of the criticism of grouping is directed against groups formed on the basis of ability Doubtless, nobody would desire a group possessing the maximum degree of heterogeneity, even in intellectual ability, and certainly not in chronological age, physical maturity, background, motivation, and the like It is probable therefore, that *everybody wants a group with a certain degree of homogeneity* The differences arise regarding the degree and basis of the homogeneity ²⁷

Arguments for and against ability grouping. An imposing list of a dozen or more arguments for, and an equal number against, ability grouping has been assembled ²⁸ The crucial point at issue is Do groups formed upon the basis of ability aid or hinder learning? Among the alleged advantages it is argued that ability grouping makes it easier to adapt instructional materials and methods to the individual pupil, thereby stimulating bright pupils and encouraging dull pupils, with the result that achievement is increased and failure reduced Among the alleged disadvantages, on the other hand it is argued that the system is essentially undemocratic and that any gains in academic achievement are likely to be slight in amount and purchased at too dear a price, since the bright pupils tend to graduate too young and to develop a sense of superiority, while dull pupils may overwork or may develop a sense of inferiority Here as always, however, it is impossible to decide a scientific question merely by counting the arguments pro and con or by attempting to weigh the logic or fervor with which they are advanced Fortunately, on this problem a considerable amount of experimental work has been done although most of the studies must be characterized as inadequate and inconclusive

The experimental evidence Summaries of the experimental literature relating to ability grouping have been made by Billett,²⁹ Wyndham,³⁰

²⁷ Cf Henry J Otto *Elementary School Organization and Administration* (Second Edition) page 184 New York D Appleton Century Company, 1944

²⁸ For rather complete summaries of the arguments see Austin H Turney, *The Status of Ability Grouping* *Educational Administration and Supervision* 17 23 January 1931 *Ninth Yearbook of the Department of Superintendence* pages 121 126 Washington D C National Education Association 1931 and Ernest W Ties *Tests and Measurements in the Improvement of Learning* pages 262-264 Boston Houghton Mifflin Company 1939

²⁹ Roy O Billett *op cit* pages 16-37

³⁰ Harold S Wyndham *Ability Grouping* page 128-159 Melbourne Australia Melbourne University Press 1934

Cornell,³¹ and various writers in the *Review of Educational Research*³² In 1931 a foreign observer³³ commented upon the 'haphazard condition' of the research upon the problem and pointed out that the experimental studies "raise more issues than they settle"

Ten years later an American educator³⁴ saw 'little or no solid, objective evidence upon which to base a decision as to the effectiveness of homogeneous grouping as actually practiced' Cornell states the situation as follows "Reviewers are generally agreed that the experimental evidence as to the achievement status of pupils under a plan of ability grouping is inconclusive"³⁵ This writer notes however that 'one of the most consistent results' has been the increased speed of progress possible by bright learners "at every level from the first grade through college," and that a reduction in the amount of failure by the less capable learners has been "rather consistently reported"³⁶ Her final conclusion is as follows³⁷

The results of ability grouping seem to depend less upon the fact of grouping itself than upon the philosophy behind the grouping the accuracy with which grouping is made for the purposes intended the differentiations in content method and speed and the technique of the teacher as well as upon more general environmental influences Experimental studies have in general been too piecemeal to afford a true evaluation of results but when attitudes methods and curricula are well adapted to further the adjustment of the school to the child results both objective and subjective seem to be favorable to the grouping

The above statement is worthy of careful study It seems reasonably clear that the evil effects of ability grouping feared by its opponents need not occur, and, on the other hand that the alluring advantages claimed by its advocates may not materialize In other words there is no money-back guarantee with ability grouping At best it merely affords more favorable conditions for doing something about the problem of individual differences The fundamental adjustments must be in terms of properly differentiated curricula and of teaching methods On this point Otto says

All authorities are agreed that no classification scheme can remove the need for adjusting instructional materials and methods to the varying needs of pupils in the group³⁸

Upon certain important issues unfortunately, there has been little or no

³¹ Ethel L. Cornell *Effects of Ability Grouping Determinable from Published Studies Thirty Fifth Yearbook of the National Society for the Study of Education Part I* pages 289-304 Bloomington Illinois Public School Publishing Company 1936 Quoted by permission of the Society

³² At three-year intervals beginning with Volume I 1931

³³ Harold S. Wyndham *op cit* page 156

³⁴ L. A. Williams *Secondary Schools for American Youth* page 230 New York American Book Company 1941

³⁵ Ethel L. Cornell *op cit* page 290 Quoted by permission of the Society

³⁶ *Ibid* pages 396-397 Quoted by permission of the Society

³⁷ *Ibid* page 304 Quoted by permission of the Society

³⁸ Henry J. Otto *op cit* page 190

experimentation No one for example, has determined the effect of various methods of adapting work to pupils of different levels of ability This is especially important, since the methods actually employed have usually been most effective for dull learners and least effective for bright learners In most cases, probably, the methods have been those which are used with ordinary heterogeneous groups, and which appear to be least appropriate to the more capable individuals

Nor has there been any convincing experimental attack to determine the effect of ability grouping upon the work habits and mental health of the pupils Such meager results as do exist are favorable Maller³⁹ found evidence that such desirable social traits as co-operation were developed better under a system of ability grouping It is a common observation that the best competition in sports, such as golf and tennis, is among those who "play about the same kind of game" Additional evidence that homogeneity is an attribute of natural social groups is afforded by the numerous studies which have shown that there is a positive correlation between friends of all ages as well as between husbands and wives, on practically all personality traits investigated⁴⁰ Partridge⁴¹ points out that several studies have revealed a greater similarity among friends in mental age than in chronological age But the main reliance so far has been upon questionnaire studies, of which the most extensive is by Sauvain⁴² One study of the attitude of 645 junior high school pupils toward ability grouping came to the conclusion that the great majority are happy and satisfied and that they accept and believe in the grouping that exists as the best situation for them⁴³ That the opinions of parents as well as of teachers were favorable to ability grouping in the cities where it was employed is indicated by the following conclusions⁴⁴

On the whole where grouping is used parents believe that children are at least as happy do better work in school and are correctly sectioned according to ability

Teachers seem to like ability grouping somewhat more than do the parents

They believe that grouping improves social attitudes leads to better work by pupils and increases the happiness of children

³⁹ Julius Bernard Maller *Coöperation and Competition* page 163 New York Bureau of Publications Teachers College Columbia University 1929

⁴⁰ Helen M Richardson Studies of Mental Resemblance between Husbands and Wives and between Friends *Psychological Bulletin* 36 104-120 February 1939

⁴¹ E DeAlton Partridge *Social Psychology of Adolescence* Chapter V New York Prentice-Hall Inc 1938

⁴² Walter Howard Sauvain *A Study of the Opinions of Certain Professional and Non Professional Groups Regarding Homogeneous or Ability Grouping* 151 pages New York Bureau of Publications Teachers College Columbia University 1934

⁴³ Austin H Turney and M F Hyde The Attitude of Junior High School Pupils toward Ability Grouping *School Review* 39 606 October 1931

⁴⁴ Walter Howard Sauvain *op cit* pages 115 116

The technique of ability grouping. There is no general agreement as to the best basis for ability grouping. In fact, there is probably no one "best basis." Much depends upon the local conditions, the data available, the nature of the subject, the size of the school, the fundamental philosophy of the school, and the like. It is often true, as one writer suggests, that "the soundest policy in dealing with educational measurements is to obtain objective data and interpret them subjectively."⁴⁵ Nor is there uniformity in either theory or practice regarding the number and size of the groups, the proper differentiation in methods and curricula, or the relative emphasis upon acceleration and enrichment for the bright groups.

A useful distinction is made between vertical and horizontal classification. Vertical classification attempts to bring together pupils of approximately the same *status*. The successive grade levels of the ordinary school represent such an attempt. The basis is usually CA, or some combination of CA, MA, and EA. The use of the average of the MA and EA, or the average G-score on an intelligence test and a general achievement test, has much to commend it in the intermediate and upper classes.⁴⁶ Horizontal classification means that on any grade level the pupils are further divided according to ability, or *rate* of learning. For this ability grouping in the academic subjects, the IQ, or a combination of IQ and CA, is probably most often employed. Boyer shows how a two-way distribution of IQ and CA, divided by horizontal and vertical lines, may be used effectively for this purpose.⁴⁷ In the high school, aptitude tests are sometimes better than general intelligence tests. In other words, the purpose is to bring together for instructional purposes those pupils who represent approximately the same educational and mental status, and who are capable of progressing in the subject at about the same rate. The system should be flexible enough to permit the shifting of pupils from one group to another in any subject whenever it is evident they are improperly classified in that subject. For non-academic subjects, such as woodwork and music, for extracurricular activities, and possibly for the homeroom in high school, the groups may be as heterogeneous as the population of the school. Such a flexible program is inherently democratic. Small schools are of necessity limited to informal groupings made within the classroom.

But the most important problem of adjustment yet remains for the classroom teacher. She must study the individual pupils in her class, whenever necessary must divide them into temporary groups for remedial instruction, and must vary the instructional materials and teaching methods as conditions seem to warrant. In the last analysis, the adjustment of the

⁴⁵ Jacob S. Orleans, *Measurement in Education*, page 286. New York: Thomas Nelson and Sons, 1937.

⁴⁶ Cf. William M. McCall, *Measurement*, Chapter XI. New York: The Macmillan Company, 1939.

⁴⁷ *Thirty-Fifth Yearbook*, *op. cit.*, pages 199-203.

school to individual differences becomes a teaching problem. As McCall says, "But after all how pupils are taught and not how they are grouped is the vital matter"⁴⁸ Turney puts the matter concisely:⁴⁹

The actual sectioning is but a minor part of ability grouping; the real job rests with the teachers. To adjust subject matter so that a child can use his mental ability and to adjust method so that he will use it—these are the outstanding problems. For it is idle to talk of effective development unless children can and do use their mental ability.

Special classes are sometimes formed for pupils at the extremes of the distribution, although in high school those for the very slow learner are much more frequent than those for the very bright.⁵⁰ In such classes the teaching is highly individualized. Patience, skill in diagnosing pupil difficulties and training in mental hygiene are important qualifications for teachers of slow classes. High intelligence, versatility, sound scholarship, and a thorough grounding in psychology are essential qualifications for teachers of special classes for bright pupils.

It is doubtless possible to overdo the idea of "special" classes and schools of one sort or another. Although in a real sense every pupil is unique and should receive special attention, it would certainly be a grave mistake to become so occupied with the "exceptional" pupils as to overlook adequate educational provision for the larger group, who are, to all intents and purposes, "perfectly normal." This situation has been satirized as follows:⁵¹

Johnny Jones has lost a leg
 Fanny's deaf and dumb,
 Marie has epileptic fits
 Tom's eyes are on the bum
 Sadie stutters when she talks
 Mabel has T.B.
 Morris is a splendid case
 Of imbecility
 Billy Brown's a truant
 And Harold is a thief
 Teddy's parents gave him dope
 And so he came to grief
 Gwendoline's a millionaire
 Gerald is a fool
 So every one of these darned kids
 Goes to a special school
 They've specially nice teachers
 And special things to wear
 And special time to play in
 And a special kind of air
 They've special lunches right in school

⁴⁸ William A. McCall *op cit* page 168

⁴⁹ *Thirty-Fifth Yearbook op cit* pages 113-115. Quoted by permission of the Society.

⁵⁰ Roy O. Billett *op cit* page 196

⁵¹ Elmer Harrison Wilds *The Foundations of Modern Education* page 523 New York Farrar & Rinehart Inc. 1942

While I—it makes me wild—
I haven't any specialties,
I'm just a normal child.

Acceleration and retardation. In the elementary school a common device for reducing the heterogeneity of the class is to eliminate the extremes of the distribution at promotion time. To do this a small number of the most capable pupils are allowed to "skip" a grade or half grade, and usually a larger number of the least capable pupils are "failed," or "retained" in the same grade for another year or half year. Witty and Wilkins⁵³ published a critical survey of the literature relating to acceleration, and, in spite of certain limitations in the studies, concluded that "most reports show clearly that acceleration, when practiced, is associated with desirable adjustment in all types of development for which data have been assembled." One experiment in which pupils allowed to skip a grade were paired with pupils of like ability not skipped led to the conclusion that "under reasonably favorable conditions skipping is a satisfactory method of accelerating pupils of superior ability."⁵⁴

Recent studies have attempted to determine the effect of acceleration upon the pupils' personality and social adjustments in high school and college, apparently accepting Terman's verdict regarding the academic achievement of superior pupils. "The earlier they enter college the better work they do there, at least down to an entrance age of 15 years."⁵⁵ Almost without exception the results appear to be favorable. Engle, for example, found that accelerated students in high school when compared with other students of their own chronological age were "at least as active socially as non-accelerated students."⁵⁶ In 1943, Pressey⁵⁶ surveyed the literature regarding acceleration on the college level and came to the conclusion that "the great majority of accelerated students do well in school, are socially adjusted, do not suffer in health, and are not handicapped in after-school career."

During World War II great emphasis was placed upon accelerated programs of education, particularly on the college level. Several colleges have attempted to investigate the effect of these programs upon the students

⁵³ Paul A. Witty and Laroy W. Wilkins, "The Status of Acceleration or Grade Skipping as an Administrative Practice," *Educational Administration and Supervision* 19: 321-346, May, 1933.

⁵⁴ Jesse E. Adams and C. C. Ross, "Is Skipping Grades a Satisfactory Method of Acceleration?" *American School Board Journal* 85: 24-25, July 1932.

⁵⁵ Lewis M. Terman, "The Gifted Student and His Academic Environment," *School and Society*, 49: 68, January 21, 1939.

⁵⁶ Thelburn L. Engle, "A Study of the Effects of School Acceleration upon the Personality and Social Adjustments of High School and University Students," *Journal of Educational Psychology* 29: 523-529, October 1938.

⁵⁷ S. L. Pressey, "Acceleration versus Lock Step," *Educational Research Bulletin*, 22: 29-30, February 17, 1941.

Studies directed by Pressey⁵⁷ at Ohio State University have been especially noteworthy. With few exceptions the results have favored acceleration. Another investigation⁵⁸ concludes that many more superior women students than usually attempt it "can complete a college program in three years or less without unfortunate effects as regards scholarship, recreation, health, or after-school career."

In a comprehensive, heavily documented monograph Pressey summarizes the literature on acceleration and calls upon educators to break the lock-step curriculum at all levels for the abler students.⁵⁹

There is evidence to suggest that superior children might best begin school somewhat younger than average children—that school entrance should be on the basis of total maturity rather than simply on arrival at the chronological age of six. In elementary and secondary schools, even though accelerates have usually not been selected carefully on any broad basis or helped to adjust to the more advanced work and new companions, a majority of them have nevertheless continued to show academic superiority and good social adjustment after acceleration. When accelerates have been carefully chosen as intellectually superior and well adjusted, means provided for facilitating advancement, as by adjustment classes, or sections arranged for superior students who then move forward together to cover perhaps three years of junior high school work in two, results have been found almost distressingly satisfactory. Such students do well in later public-school work (as in senior high school) and in their relationships with pupils in the classes into which they have been advanced, sometimes even better than students of equal ability and academic record at the time the special program of acceleration began who continued at the regular pace. *The conclusion seems unavoidable that the usual lock step grossly wastes the time of the ablest young persons.* In college students who entered early as a result of acceleration similarly show the highest academic record, the lowest academic mortality, and high participation in campus life.

The weight, both of the arguments and of experimental evidence, appears to be against failure or retardation as a school policy. In Otto's⁶⁰ survey the literature indicated that about 20 per cent of repeaters do better and 40 per cent do worse than before. He concluded that if the objective of the modern school is the optimum development of its pupils, "non-promotion is not the way to get it." Several studies indicate the value of trial promotions. An investigation by McKinney,⁶¹ for example, involving more than 13,000 pupils, shows a saving of about three out of every four

⁵⁷ Cf. S. L. Pressey and S. B. Folk, "First Evaluations of an Accelerated Program in a College of Engineering," *Journal of Engineering Education* 34: 477-480, March 1944; S. L. Pressey, "Acceleration the Hard Way," *Journal of Educational Research* 37: 561-570, April 1944.

⁵⁸ Marie A. Flesher, "An Intensive Study of Seventy-Six Women Who Obtained Their Undergraduate Degrees in Three Years or Less," *Journal of Educational Research*, 39: 602-612, April, 1946.

⁵⁹ Sidney L. Pressey, *Educational Acceleration, Appraisals and Basic Problems* (Bureau of Educational Research Monograph No. 31), page 143, Columbus, Ohio: Ohio State University, 1949. Italics added.

⁶⁰ Henry J. Otto, *op cit*, page 232.

⁶¹ H. T. McKinney, *Promotion of Pupils: A Problem in Educational Administration*, 206 pages, Urbana, Illinois: University of Illinois, 1921.

repeaters. One study has shown that the threat of failure affords ineffective motivation.⁶² Certainly, with a modern curriculum and an adequate program of diagnosis and guidance few if any failures should occur.

Continuous promotion. Otto⁶³ has proposed a somewhat theoretical but very suggestive promotion plan for the elementary school, which abolishes not only acceleration and nonpromotion but the term 'school grade' as well. Such a type of organization has been in successful operation in several school systems for a number of years.

His plan involves the following five essential features:

1. There would be available extensive data of an objective character on each child, so that he may be placed at all times in groups in which he can work to the best advantage in terms of his own developmental readiness.

2. There would be continuous pupil adjustment and progress with shifts from one group to another at any time during the year that a change would seem advisable.

3. The major classifications which take place in the ordinary school at the beginning of each term would be eliminated.

4. It would make possible longer teacher-group relationships in which 'the same teacher works with the same group of children for two or three consecutive semesters or years.'

5. The conventional competitive marking system would be replaced with 'extensive, objective cumulative data on many aspects of the growth and development of each child.'

SELECTED REFERENCES FOR FURTHER READING

- Berkshire James R. Improvement in Grading Practices for Air Training Command Schools. *ATRC Manual 50-900-9 Headquarters Air Training Command*
Scott Air Force Base Illinois June 1952 39 pages
- Cook Walter W. The Functions of Measurement in the Facilitation of Learning Chapter 1 in F. F. Lindquist (Editor) *Educational Measurement* Washington D. C. American Council on Education 1951
- Crow Lester D. and Crow Alice. *Educational Psychology* New York American Book Company 1948 Chapters 11 and 21 Educational Implications of Individual Differences and Adjustment of Exceptional Individuals
- Educational Policies Commission. *Education of the Gifted* Washington D. C. National Education Association 1950 88 pages
- Lewis Robert S. *Educational Psychology a Problem Approach* New York D. Van Nostrand Company 1951 Chapters II, XII and XV The Admission and Classification of Pupils Their Achievement and Their Elimination from School
- Classroom Tests and Marks and Exceptional Children
- Gilbert Arthur W. Design and Pattern of the Curriculum. *Review of Educational Research* 21 196-208 June 1951

⁶² Henry J. Otto and Ernest O. Melby. A Attempt to Evaluate the Threat of Failure as a Factor in Achievement. *Elementary School Journal* 35 588-596 April 1935

⁶³ Henry J. Otto op cit 236-242

- Hildreth, Gertrude H, *Educating Gifted Children at Hunter College Elementary School* New York Harper & Brothers, 1952 272 pages
- Hollingshead, Augustus B, *Elmtown's Youth, the Impact of Social Classes on Adolescents* New York John Wiley & Sons, 1949 Chapters 8 and 13, "The High School in Action" and "Leaving School "
- Mackenzie, Gordon N and Bebell, Clifford, "Curriculum Development," *Review of Educational Research*, 21 227-237, June, 1951
- Pressey, Sidney L, *Educational Acceleration* (Bureau of Educational Research Monograph No 31) Columbus Ohio Ohio State University, 1949 153 pages
- Terman, Lewis M, and Oden, Melita H, *The Gifted Child Grows Up* Stanford, California Stanford University Press, 1947 148 pages
- Witty, Paul (Editor), *The Gifted Child* Boston D C Heath and Company, 1951. 338 pages

14

Evaluation in Guidance

The field of guidance developed rapidly from 1918 onward, concurrently with the beginnings of mass group testing. Now a vast body of information and techniques must be mastered before one can qualify as a competent professional guidance worker, so it is not feasible in an elementary measurement textbook to do more than make a few brief remarks concerning evaluation in school guidance programs. The interested student will want to scan the selected references at the end of this chapter for supplementary reading material.

A. The Meaning and Importance of Guidance

The fundamental problem of life is adjustment. At birth the human infant is much less well adjusted to the world in which he must live than many of the simpler organisms. Man's dominant place in the universe is due largely to his remarkable capacity for modifying his reactions in the direction of a more adequate adaptation to the conditions under which he must live. The process by which these changes take place is called *learning*, and the result is called *education*. The function of the school is to provide a favorable environment in which these changes may take place. The role of the classroom teachers and of the school administrators is to stimulate and to direct the learning process.

The aim of all guidance is to assist the learner to acquire sufficient understanding of himself and of his environment to be able to utilize most intelligently the educational opportunities afforded by the school and the community. The problem of guidance arises from the fact that an immature but growing individual with a unique combination of abilities and limitations is confronted with a complex and ever changing environment. Guid-

ance used to be regarded as an effort "to see through Johnny and to see Johnny through." The emphasis today has shifted to an effort "to help Johnny see through himself and to see himself through."¹ It seeks to assist each student to choose, and make satisfactory progress in, those activities which will contribute most to his development, individual happiness, and social worth.

Certain circumstances have conspired to make guidance one of the most important responsibilities of the modern school. This is particularly true of the secondary school and of the college. Figure 12 on page 248 shows that in 1900 4 per cent of all students in the United States were enrolled in secondary schools and 1.4 per cent in institutions of higher learning while in 1950 the corresponding percentages were 18.1 and 6.2. During this same period the total number of students doubled, rising from 17,200,000 to 34,400,000. As a result of these conditions the student body of the modern secondary school and college represents a greater diversity of backgrounds, interests, ambitions, and abilities than has ever been true before.

At the same time science and invention have greatly complicated and are constantly changing the social and economic world from which these pupils come and to which they must return. Likewise, the school situation itself academically as well as socially has markedly increased in complexity. The small high school with a single curriculum leading to college has tended to give way to larger schools with a more diversified program. Judd² called attention to the fact that the number of subjects offered in American high schools increased from 9 in 1890 to more than 250 in 1942. At the present time a pupil of high school age in a large modern American city has a choice of a score or more different curricula.

As it is always easier for the traveler to lose his way in a large city than in a small town, especially if he is lacking in maturity and experience, it is perhaps not surprising that many of those who enter the modern secondary school and college never succeed in making a satisfactory adjustment to these institutions. Likewise the vast number of adolescents and adults who find their way into penal institutions or into hospitals for the physically and the mentally ill, and the much larger number of others who lead unhappy and unsuccessful lives, afford tragic evidence that adjustment outside the school has been equally unsatisfactory. There seems no escaping the fact that when the conditions of life increase in complexity, the need for guidance increases proportionately. The better the guidance program, the less will be the need for diagnostic and remedial work later on. An adequate guidance program is the best form of prevention.

¹ George E. Myers, *Principles and Techniques of Vocational Guidance*, page 4, New York: McGraw-Hill Book Company, 1941.

² Charles H. Judd, *General Education and the Baccalaureate Degree*, *School and Society*, 56, 30 July 11, 1942.

B. The Place of Measurement in Guidance

Two errors are common in assigning the place of measurement in guidance. The first of these, fortunately less common now than in the early days of standard testing, is to think of guidance as synonymous with testing. *Guidance is always more than the giving of tests, no matter how extensively or carefully done.* As a matter of fact, whether or not tests serve any guidance function depends upon the use made of the results. Here, as elsewhere, tests are merely tools. The second error, very common today, is to dismiss measurement altogether and to regard it as wholly unessential to guidance if not indeed an actual obstacle. This viewpoint is as extreme as the first, however. While testing is never everything in guidance, it is almost always something. In fact, it may be asserted confidently that *evaluation in some form is implicit in the guidance function.* Properly used tests are valuable aids to self-analysis.

Fowler¹ indicates that many common mistakes will be avoided if users of tests in guidance will remember at all times that "the only justifiable reason for using tests in the guidance program is to serve the individual inventory in counseling." He formulates seven "guiding rules" based upon this point of view.

1 Any item of the individual inventory, whether it be a test score, a teacher mark, a fact about the pupil's health, can be interpreted in the counseling situation only in the light of all the other inventory data having some bearing on the problem at hand. This is to say, a chief value of test scores is the check which they provide upon the meaning of other accumulated facts. In turn, the importance to be accorded test scores in any given case must be weighed in the light of other data from the individual inventory. Dependence must be placed upon tests to supply facts when they have not been accumulated through other means.

2 Test scores, like other items in the inventory, must be interpreted cautiously until norms are scientifically established for the local situation and for the particular kind of problem which the pupil presents.

3 The meaning of a test score may not be the same from one pupil to another because of the differences in other pertinent inventory data. The meaning may change even for the same pupil from one problem to another or from one time to another.

4 Real counseling will encourage decisions or judgments only on the basis of as full an inventory of pertinent facts as possible. Thus several measures are usually better than just one or two. Likewise, the same dependence will not be placed upon so-called 'interest' or 'personality' tests as upon achievement and aptitude tests.

5 It is recognized that certain tests are regularly used in the school by the administrator in pupil classification and curriculum planning. They are used by teachers in individualizing teaching methods. The data from these same tests are of even greater use for counseling and should always be recorded in the cumulative record. Tests used by the administrator for these purposes may supplement the

¹ Fred M. Fowler, "To Inquirers about Tests," *Education for Victory* 3:12-13, December 4, 1944.

tests used only by the counselor. This fact should not be overlooked in their choosing.

6 Tests are best used as aids to *counseling* rather than as standards for arbitrary selection (or rejection) for training and job opportunities.

7 Familiarity with a test gained through its use, is important. In deciding to use a new test to measure the same traits, loss of this familiarity should be weighed carefully against the possible gain in reliability, validity, usability, and economy.

If these suggestions are kept in mind, the dangers against which Rogers⁴ warns will be avoided.

Counseling is the process of assisting the individual in making the maximum adjustment to the educational opportunities of his environment in terms of his abilities, interests and needs. It is normally a face-to-face relation between an older, more experienced person and a less mature person. It is an example of co-operative problem solving. The role of the counselor is not to make decisions for the pupil, but rather to help him to solve his own problems intelligently. The pupil's part naturally increases with his maturity, the ultimate purpose of counseling being so to develop the pupil's self reliance that outside help becomes progressively unnecessary.

Counseling is not a new development in education. In some type or other it has probably existed longer than formal education itself. The difficulty has been not with the amount of counseling available but with its quality. Nowhere is the wisdom of the homely American philosopher, Josh Billings, more apparent than in counseling: "It is better to know less, than to know so much that ain't so." The improvement of measurement techniques in the present century and the development of cumulative record forms have made it possible to substitute factual data for opinion and hearsay.

C Guidance Is a Co-operative Venture

In a very real way all persons in contact with the child—teachers, administrators, specialized educational personnel, clinic staffs, parents, ministers, law enforcement officials, and others in the community—are guidance workers. They must work together effectively in order to prevent juvenile delinquency, inadequate benefit from schooling, vocational frustration, neurotic inadequacy, and mental illness. This calls for continuous co-ordinated effort by everyone. No one person or agency can shoulder the burden alone, for no single group has the necessary training, experience and facilities. Legal, medical, social work, economic, and political considerations often loom fully as important as purely school based educational aspects. The teacher should remember eternally that his is a co-operative task where guidance is concerned. Though it is undoubtedly true in a certain sense that "teaching is guidance," much more guidance than can be

⁴ Carl R. Rogers, "Psychometric Tests and Client-Centered Counseling," *Educational and Psychological Measurement* 6: 139-144, Spring, 1946.

given in the classroom under present conditions is essential if America's children are to develop optimally

SELECTED REFERENCES FOR FURTHER READING

- Arbuckle, Dugald S., *Teacher Counseling* Cambridge, Massachusetts Addison Wesley Press, 1950 179 pages
- Bennett, George K., and Seashore, Harold, "Testing for Vocational Guidance in High School," pages 71-79 in Henry Chauncey (Chairman) 1947 Invitational Conference on Testing Problems, "Exploring Individual Differences" American Council on Education Studies, Series I, No. 32, Vol. XII, October, 1948
- Bennett, George K., Seashore, Harold G., and Wesman, Alexander G., *A Manual for the Differential Aptitude Tests* (Second Edition) New York Psychological Corporation, 1952 77 pages
- Bledsoe, Joseph C., "Success of Non High School Graduate GFD Students in Three Southern Colleges," *College and University* 29 381-388, April, 1953
- Coleman, William, and Cobb, E. B., *Guidance Use of Test Results* Professional Manual No. 2, Tennessee State Testing Program Knoxville, Tennessee University of Tennessee, November, 1951 47 pages
- Crawford Albert Beecher, and Burnham, Paul Silvester, *Forecasting College Achievement A Survey of Aptitude Tests for Higher Education*, Part I New Haven Yale University Press 1946 291 pages
- Darley, John G., and Anderson, Gordon V., "The Functions of Measurement in Counseling," Chapter 3 in E. F. Lindquist (Editor), *Educational Measurement* Washington, D. C. American Council on Education, 1951
- Davis, Frederick B., *Utilizing Human Talent* Washington, D. C. American Council on Education 1947 85 pages
- Davis, Frederick B. (Editor), "Educational and Psychological Testing" *Review of Educational Research*, 23 1-110, February 1953 See reviews of the 1949-52 literature concerning tests of general mental ability (Julian C. Stanley) tests of special aptitude (William G. Mollenkopf), nonprojective tests of personality and interest (David V. Tiedeman and Kenneth W. Wilson), and projective tests of personality (John W. M. Rothney and Robert A. Heimann)
- Erickson, Clifford E., *The Counseling Interview* New York Prentice-Hall, Inc., 1950 174 pages
- Frederiksen, Norman, and Schrader, W. B., "The ACE Psychological Examination and High School Standing as Predictors of College Success" *Journal of Applied Psychology*, 36 261-265, August, 1952
- Froelich, Clifford P., *Guidance Services in Smaller Schools* New York McGraw-Hill Book Company, 1950 352 pages
- Froelich, Clifford P., and Darley, John G., *Studying Students Guidance Methods for Individual Analysis* Chicago Science Research Associates 1952 411 pages
- Lefever, D. Welty, Turrell, Archie M., and Weitzel Henry I., *Principles and Techniques of Guidance* New York Ronald Press, 1950 577 pages
- Rogers, Carl R., *Client-Centered Therapy, Its Current Practice, Implications, and Theory* Boston Houghton Mifflin Company, 1951 560 pages
- Rothney, John W. M., and Roens Bert A., *Guidance of American Youth an Experimental Study* Cambridge, Massachusetts Harvard University Press, 1950 269 pages.

- Rothney, John W. M., and Danielson, Paul J., 'Counseling,' *Review of Educational Research*, 21 132-139 April, 1951
- Rothney, John W. M., "Interpreting Test Scores to Counselees," *Occupations* 30 320-322, February, 1952
- Strang, Ruth, 'Major Limitations in Current Evaluation Studies,' *Educational and Psychological Measurement* 10 531-536 Autumn (Part 2), 1950
- Strang, Ruth, '7 Ways to Improve the Rating Process,' *Occupations*, 29 107-110, November, 1950
- Super, Donald F., *Appraising Vocational Fitness by Means of Psychological Tests* New York Harper & Brothers 1949 727 pages
- Traxler, Arthur F., and Townsend, Agatha (Editors), *Improving Transition from School to College* New York Harper and Brothers, 1953 165 pages
- Willey, Roy D., *Guidance in Elementary Education* New York Harper and Brothers, 1951 825 pages
- Williamson, E. G. and Foley, J. D. *Counseling and Discipline* New York McGraw Hill Book Company 1949 385 pages
- Wolfe, Dael, and Oxtoby, Toby, 'Distributions of Ability of Students Specializing in Different Fields,' *Science* 116 311-314, September 26, 1952.

15

Evaluation of Schools

A. The Problem of Evaluation

The motivational effects of evaluation programs have long been evident in the schools. If achievement is rated by tests, both teachers and pupils work to pass the tests. If progress is appraised in other ways, activities related to these methods of evaluation are evident in the daily or weekly school program. The modern point of view is that evaluation is not a series of periodic examinations applied externally but an intrinsic part of the learning process with its planning, evaluating cycle. Tests do, of course, have their places in this process. Viewed thus, methods of evaluation can be one of the most valuable tools for creating interest and purpose in further learning.¹

Measurement and evaluation. As used in education, *evaluation* is a far more inclusive concept than *measurement*. Two aspects of evaluation may be distinguished: (1) data relating to some important aspect of the school, such as its organization, program, or results, and (2) a set of values or standards against which these data are interpreted and appraised. Furthermore, the evaluator's educational philosophy and sense of values will determine what objectives of the school program he considers to be important, as well as what data he will look for, or regard as relevant in the situation. It is apparent that while measurement may be highly mechanical and at times a routine, evaluation can never be, at every stage evaluation requires the exercise of mature judgment.

Measurement implies the use of some tool or instrument, such as a test or scale, and provides a quantitative description of observed phenomena.

¹ Ernest R. Hilgard and David H. Russell, 'Motivation in School Learning' page 64 in the *Forty Ninth Yearbook of the National Society for the Study of Education Part I (Learning and Instruction)* Chicago: University of Chicago Press 1950. Quoted by permission of the Society.

This is always desirable but it should never exclude relevant data of a subjective and qualitative character, or the consideration of outcomes not immediately observable. Some writers have criticized existing measurement in education for the reason that it furnished inadequate data for evaluation.² At best measurement merely provides data needed for evaluation; it is not evaluation per se.

The *Sixteenth Yearbook*³ of the Department of Elementary School Principals of the National Education Association presents a good discussion of evaluation on the elementary level. The ten chapters of this report are given below:

- I The Fundamentals of School Appraisal
- II Appraising the School Organization
- III Appraising Administrative and Supervisory Procedures
- IV Evaluating the Curriculum
- V Appraising Methods of Learning and Teaching
- VI Evaluating Socializing Experiences
- VII Measuring the Progress of Pupils
- VIII Estimating the Efficiency of Teachers
- IX Judging School Equipment
- X A Review of the Technics of Appraisal

Note that the term "measuring" occurs in only one chapter heading. The other terms—appraising, "evaluating," "estimating," and "judging"—all similar in meaning, imply the use of techniques that go beyond testing and examining.

The Eight Year Study of the Progressive Education Association,⁴ which included both elementary and secondary education, and the Three-Year Study of the Commission on Teacher Education⁵ on the college level are illustrations of an enlarged conception of evaluation. The committee sought to devise suitable instruments of measurement for outcomes—such as interests, attitudes, creativeness, and various aspects of thinking—less tangible than those measured by ordinary tests and examinations. It also utilized other types of data, such as anecdotal records, family histories, records of the pupils' activities, and the like.

Possibly no more ambitious example of this enlarged conception of evaluation is available than the Cooperative Study of Secondary School Stand-

² Cf. Verner M. Sums, "Educational Measurement and Evaluation," *Journal of Educational Research* 38: 18-33, September 1944.

³ "Appraising the Elementary-School Program," *The National Elementary Principal* 16: 227-600, July 1937.

⁴ Eugene R. Smith, Ralph W. Tyler, and Evaluation Staff, *Appraising and Recording Student Progress*, 500 pages, New York: Harper & Brothers, 1942.

⁵ Maurice E. Troyer and C. Robert Pace, *Evaluation in Teacher Education*, 368 pages, Washington: American Council on Education, 1944.

ards * Table 37 shows the six main methods the study employed, of which only one represents the use of measurement in the ordinary sense

TABLE 37

THE MAIN METHODS OF EVALUATION USED BY THE COOPERATIVE STUDY OF
SECONDARY SCHOOL STANDARDS WITH THE WEIGHT ASSIGNED TO EACH

Method		Per Cent
1	Evaluative Criteria	40
A	Educational Program	20
	Curriculum	28
	Pupil activity	28
	Library	28
	Guidance	28
	Instruction	60
	Outcomes	28
B	Organization and Plant	20
	Staff	100
	Administration	60
	Plant	40
2	General Judgment* by Visiting Committees	20
3	Growth as Measured by Standard Tests	20
4	Success of Pupils	10
A	In College	10 to 1
B	Noncollege	0 to 9
5	Judgment by Pupils	6
6	Judgment by Parents	4
Total		100

The importance of evaluation. Without some form of evaluation everything about education becomes a matter of blindly hoping that all is well. In the critical period shortly before the Civil War, Abraham Lincoln began an important address with this statement: "If we could first know where we are and whither we are tending, we could better judge what to do and how to do it." It is no less true in education than in government that we must first "know where we are," and especially "whither we are tending," before we are in a position to judge intelligently regarding "what to do and how to do it." In the final analysis it is the function of all attempts at evaluation to afford a basis for rational action. Apparently, educators have always recognized this, even if often somewhat vaguely. For example,

* For a full account of this study see *Evaluation of Secondary Schools General Report* 526 pages Washington D. C. Cooperative Study of Secondary School Standards 1939. For a briefer statement see *How to Evaluate a Secondary School* (1940 Edition) 139 pages Washington D. C. Cooperative Study of Secondary School Standards.

a college president said "Self-criticism and self-appraisal (now a 'self survey' or an 'evaluation') are as old as education" ⁷

The emphasis today is more and more upon the importance of self-evaluation. This holds true for all levels of education from the activity of the pupils in an elementary class passing judgment upon the success of a unit of instruction planned and executed by themselves to the formal Report of the Harvard Committee ⁸

Many years ago Thorndike pointed out that "the actual changes wrought in boys and girls by this or that form of education are being measured, old and new methods are being tested by experiment in the same spirit of zeal and care for the truth that animates the man of science, and the educational customs which have been accepted unthinkingly by 'use and wont' are being required to justify themselves to reason" ⁹ Although it is probably true that more progress has been made in that direction since the statement above was written than in all the centuries preceding, improvements in evaluation procedures have hardly kept up with those in curricula and teaching methods ¹⁰

The difficulty of evaluation. The problem of evaluating education is immensely complicated. Many approaches toward a solution have been made, and none has been entirely satisfactory. For many years the various regional associations attempted to evaluate the secondary schools and colleges of America indirectly by their *possessions* rather than directly by their *products*. Such measures as the size and qualifications of the staff and the number of books in the library were at best indications of *educational opportunity*, and even to a less extent were such things as the number or type of buildings in the school plant and the amount of financial support available. The limitations of such a procedure have been characterized as follows: The standards used were mechanical, rather than vital, rigid, rather than flexible, deadening, rather than stimulating, traditional, rather than progressive, academic, rather than liberal, broadly comprehensive and subjective, rather than scientific ¹¹ Intensive and extensive study of the problem by committees within the past decade has increasingly revealed its complexity. The Cooperative Study of Secondary School Standards, for example, extended over a period of six years and cost about a quarter of a million dollars. It employed the six major methods of evaluation given in

⁷ Henry M. Wriston "A Critical Appraisal of Experiments in General Education" *Thirty-Eighth Yearbook of the National Society for the Study of Education Part II*, page 303. Bloomington, Illinois: Public School Publishing Company, 1939. Quoted by permission of the Society.

⁸ *General Education in a Free Society*. 267 pages. Cambridge, Massachusetts: Harvard University Press, 1945.

⁹ Edward L. Thorndike *Education: A First Book*, pages 7-8. New York: The Macmillan Company, 1912.

¹⁰ Cf. Pedro D. Orata "Evaluating Evaluation" *Journal of Educational Research* 33: 641-661, May 1940.

¹¹ *Evaluation of Secondary Schools: General Report on the* pages 33-50.

Table 37, and developed three scales, whose composition is given in Table 38. It will be noted that the complete scale, Alpha, includes 110 different

TABLE 38

COMPOSITION OF 1940 EDITION OF THE ALPHA, BETA AND GAMMA SCALES FOR EVALUATING SECONDARY SCHOOLS

Area	Number of Thermometers		
	Alpha	Beta	Gamma
Curriculum and course of study	19	8	3
Pupil activity program	13	5	2
Library service	11	7	3
Guidance service	7	3	2
Instruction	6	3	2
Outcomes	18	7	2
School staff	18	9	6
School plant	11	4	3
School administration	7	4	2
Total	110	50	25

"thermometers," all relating to the nine evaluative criteria of the first method listed in Table 38. In 1942 Eulich, Pace, and Ziegfeld¹² surveyed the literature of the field and came to this conclusion: "No simple and inexpensive technique has as yet been devised nor is one likely to be devised that will provide an evaluation of an entire educational program."

Evaluating teaching efficiency. A single illustration will show the complexity of the problem of evaluation. How can one best judge the worth of any particular classroom teacher? This is manifestly an important question. To a large extent the selection, growth, and promotion of teachers depend upon the answer. In general, the methods used are of three types. In the first place, tests and rating scales have been devised for measuring the *personality* of the teacher. As a rule, these have proved disappointing. The difficulty with this approach has been clearly pointed out by McCall: "No one has demonstrated just what causal relationship, if any, exists between possession of these various attributes and desirable changes in pupils."¹³

A second method attempts to measure the worth of the teacher by her *performance*, usually her activity before the class. For this purpose various score cards and rating scales have been devised. In fact, the typical rating scale attempts to secure various measures of the teacher's performance together with measures of certain traits of personality deemed important.

¹² Alvin C. Eulich, C. Robert Pace, and Edwin Ziegfeld, "Evaluative Studies," *Journal of Educational Research*, 12: 521-533, December, 1912.

¹³ William A. McCall, *Measurement*, page 403, New York: The Macmillan Company, 1939.

in teaching. But except as instruments of self-analysis by the teachers themselves, the practical value of rating scales is slight. For example, when the gains on the Stanford Achievement Test from November to May by pupils in four Wisconsin schools were used as a criterion, the correlations with 17 of the best-known measures of teaching ability available, although somewhat inconsistent, with few exceptions were so low that they could reasonably be supposed to have arisen from a population in which the true relationships were zero.¹⁴

A third method has been the attempt to judge the worth of the teacher by her *product*, the performance of her pupils. This is certainly the most direct, and is often asserted to be the only valid, approach. The most obvious way to achieve this result is to measure the improvement made by the pupils during a period of instruction under the teacher. But the problem is far more complicated than it at first appears. Even when allowances are made for differences in the intelligence and initial achievement of the pupils, the greater problem remains of determining how much of the growth is due to natural maturity and how much to the total educational environment in school and out of school, and the still greater problem of knowing how much of this improvement is due to the influence of any particular teacher. Most competent observers today would agree with Traxler¹⁵ that "the use of test results for rating teachers is seldom advisable."

A summary by Barr¹⁶ notes encouraging progress but emphasizes that the road ahead is long and difficult:

The influence of any particular teacher is deeply enmeshed in a host of other school, pupil, and community factors. While very definite progress has been made in this area, it is not easy to isolate the effects of particular teachers in particular situations. There is reason to be optimistic about the use of more precise instruments of measurement in the management of the teaching personnel, but for the time being, discretion is the best part of valor.

The Cooperative Study. One of the most ambitious attempts at evaluation by means of standard tests has been the Cooperative Study of Secondary School Standards, involving 198 schools and a total of over 300,000 tests.¹⁷ In spite of unusual care to avoid the difficulties summarized in Table 39, the Cooperative Study concluded that since the results showed that better methods of evaluation were available for accreditation, the use of standard tests should be restricted to diagnostic and guidance purposes by the local school. The Cooperative Study also attempted to judge the product of the school by follow-up studies of the subsequent careers and

¹⁴ Helen M. Walker (Editor), *The Measurement of Teaching Efficiency*, pages 73-141. New York: The Macmillan Company, 1935.

¹⁵ Arthur E. Traxler, *Techniques of Guidance*, page 186. New York: Harper & Brothers, 1945.

¹⁶ A. S. Barr, "The Use of Measurement in the Management of Teacher Personnel," *Education*, 66: 431-435, March 1946.

¹⁷ *Evaluation of Secondary Schools: General Report, op. cit.*, Chapter VIII.

TABLE 39

USE OF TESTS IN EVALUATING SCHOOLS¹²

<i>Criticism</i>	<i>Comment</i>
1 The ability of students varies widely in different schools ¹³	Each pupil's achievement can be compared with others of the same level of scholastic ability, and not with the usual national norms
2 A general testing program must be uniform and inflexible and does not recognize individual differences in schools	Difficult to meet this objection. However, there is a large body of instructional material common to all institutions. This common core can be used as a basis of comparison by testing, leaving the differentiating phrases to other types of evaluation
3 A general testing program tends to crystallize curricula and to reduce instruction to mere coaching for examinations	Danger is real, if tests are used for this purpose at regular intervals announced in advance. Does not hold for occasional testing to be used with other criteria
4 Available tests do not adequately measure important outcomes of instruction	Undoubtedly true. But less true than formerly. Moreover, many important outcomes are measured by the better tests and others are difficult to evaluate by any techniques so far developed
5 The achievement of students at any given time is the result of all previous schooling, not merely that of the present school	Can be met by using equivalent tests before and after a period of instruction, and by judging the worth of the school in terms of the changes effected between tests
6 The average achievement of an institution as a whole does not properly take into account differences within the institution in curricula and departments	A valid objection to any single criterion used for evaluating a school, and no more true of tests than of any other measure. Apparently the only answer is to present a picture of the profile of the institution showing the strong and weak points
7 Measuring the status of the pupils at any given time tends to reflect the quality of the school at some time in the past, whereas what is wanted is a picture of the school as it is now functioning	Can be met at least in part by measuring the growth produced during an instructional period in the school being evaluated
8 Some measurable outcomes may be due to out-of-school contacts and so cannot properly be attributed to the influence of the school	The objection probably holds mainly to the field of the social studies. In a sense the degree to which this occurs is a measure of the success of the school which should attempt to utilize and coordinate all instructional agencies that make up the educational environment out of school as well as in school
9 It is difficult to obtain standard conditions for administering a test to a variety of schools scattered over a wide area ¹⁴	The difficulty can be reduced to a minimum by employing a small staff of carefully trained examiners who follow a simple program fully worked out in advance

¹² Adapted largely from *Liaison of Secondary Schools**General Report Chapter VIII*

success both academic and nonacademic, of former pupils, and concluded that this method was mainly of value for local school use. A periodic canvass of the opinion of pupils about the instruction and other aspects of the school which they are attending is also a valuable means of self analysis and guidance for, although the customer may not always be right, *what he thinks* about the institution is important even when he is mistaken. In the Cooperative Study pupil judgment also proved to be about as useful for evaluating schools as the elaborate testing program.

B General Principles of Evaluation

For elementary schools. The Research Division of the National Education Association formulated the following statements of guiding principles for evaluating the programs of elementary schools, most of these appear equally applicable to the other levels of education.¹⁹

1 Adequate appraisal of the school includes more than the usual program of achievement testing.

2 School appraisal should be diagnostic, that is it should reveal the specific points of strength and of weakness in the school program.

3 Every aspect of the school program should be appraised regardless of its relative difficulty.

4 Principals and teachers should play important parts in the appraisal of their own schools. Their responsibility for planning and initiating appraisal measures will vary according to the plan of organization and administration in the school system as a whole.

5 Within reasonable limits and under proper safeguards pupils also may contribute to school appraisal.

6 Evaluation of the school program should be carried on continuously. Pertinent information should be collected thruout each year and summarized at least once a year.

7 Methods of appraisal should be selected on the basis of their reliability, practicability, and appropriateness in the particular situation to be appraised. A combination of several methods is often better than one alone.

8 Before undertaking an appraisal principals and teachers should find out how competent workers elsewhere have evaluated similar elements of the school program.

9 Careful subjective judgments formed in the light of valid criteria are better than conclusions based on objective data from a poorly planned or carelessly executed experiment.

10 Every appraisal should be made with reference to specified criteria of some kind. Such criteria should themselves be carefully evaluated before they are used.

11 Of the several types of criteria which may be used those concerned with pupil development should receive first consideration.

12 There should be close agreement between the accepted objectives of a school and the instruments which the school uses to measure its attainment of these objectives.

13 When it is impracticable to determine the merits of local school practices directly, these practices should be appraised with reference to the findings of available research studies and expert opinion outside the school.

¹⁹ *The National Elementary Principals Association* 16 237 238 July 1937

14 Statistical techniques for determining the reliability of experimental results should be used only with a thoro understanding of their purpose and significance

15 The results of appraisal should be used to improve the school program. It is essential that classroom teachers, as well as principals, be fully informed of these results

16 Parents and pupils also should be given accurate information concerning the strong and weak elements of the school program so that they may help to improve it

For secondary schools. The Cooperative Study of Secondary School Standards has prepared the following eighteen principles, which provide a comprehensive philosophy not only for evaluating secondary schools, but also for evaluating other levels of education ²⁰

1 American secondary schools much as they may differ in details are essentially alike in their underlying purposes and organization

2 In a democracy the fundamental doctrine of individual differences is as valid for schools as for individuals. Schools, as well as pupils differ from each other markedly

3 A school can be studied satisfactorily and judged fairly only in terms of its own philosophy of education its individually expressed purposes and objectives the nature of the pupils with whom it has to deal, the needs of the community which it serves and the nature of the American democracy of which it is a part. All American schools however they may differ in type have this in common they are instrumentalities for transmitting our American heritage and our American democratic ideals. Provided this aim can be clearly kept in view in every case each school is free to determine its own educational policies in promoting the ideals of American civilization

4 A school should be judged in terms of the extent to which it meets satisfactorily the needs of all pupils who should come to it not alone of those who continue their formal education in institutions of higher learning

5 Methods of accreditation and interpretation of evaluation should recognize the differences in background development and existing conditions in different states and regions. No attempt should be made to develop uniform standards for the nation or to have them administered from a single national office

6 It is more significant to measure what the school does than what it is or what it has. The educational process and product are more important to evaluate than the machinery and equipment

7 A school should be judged as a whole not merely as the sum of its separate parts

8 The number of factors evaluated in the modern secondary school should be sufficiently large and varied to give valid evidence of the worth of the school in each of its main areas

9 Accrediting criteria and procedures should be brief enough in extent sufficiently varied in form and convenient enough in arrangement to be practicable for use in all secondary schools

10 Methods of evaluation and accreditation as far as possible should be based upon scientific studies and objective evidence, rather than upon untested assumptions and unsupported opinions

²⁰ *Evaluation of Secondary Schools General Report on it* pages 57-61 also *How to Evaluate a Secondary School* (1940 Edition) *op cit* pages 17-21

11 The considered judgment of competent educators is an essential factor in the evaluation of the quality and character of the work of a school

12 A valid method of evaluation and accreditation, based tentatively upon existing research studies and expert judgment, should be fully tested by extensive experimental try-out in a large group of typical, representative secondary schools throughout the country. The results of this experimentation should be carefully analyzed and evaluated

13 While it is desirable in many respects that definite standards or levels of achievement should be developed, it is recognized that in most of the important aspects of a school's work the best available basis for the development of useful standards will probably be comparison with the practices in other comparable schools

14 A good school is a growing school. It should be judged by its progress between two different dates as well as by its status at a single date

15 Any useful, stimulating and valid method of accreditation should be flexible with the passage of time, that is, it should be capable of reasonable modification as new bases of evaluation and different levels of achievement are suggested or developed from the use of existing ones

16 If criteria for evaluation are sufficiently flexible, extensive, and thorough, it is not essential that they be applied annually

17 The bases and methods of evaluation should be such as to require active participation in the process on the part of the entire professional and non-professional staffs of the school

18 An important function of a national, regional, or state agency should be stimulation toward continuous growth and improvement, not merely inspection and admission to membership

For higher institutions. The Committee on Revision of Standards created by the Commission on Higher Institutions of the North Central Association of Colleges and Secondary Schools spent five years and \$135 000 in making a study reported in a series of seven monographs²¹ Section VI, entitled "Institutional Purposes and Clientele," and regarded as "the very heart of the new accrediting policy," is as follows²²

Recognition will be given to the fact that the purposes of higher education are varied and that a particular institution may devote itself to a limited group of objectives and ignore others, except that no institution will be accredited that does not offer minimal facilities for general education or require the completion of general education for admission

Every institution that applies for accreditation will offer a definition of its purposes that will include the following items

- 1 A statement of its objectives, if any, in general education
- 2 A statement of the occupational objectives, if any, for which it offers training
- 3 A statement of its objectives in individual development of students, including health and physical competence

This statement of purposes must be accompanied by a statement of the institution's clientele showing the geographical area, the governmental unit, or the religious groups from which it draws students and from which financial support is derived

²¹ *The Evaluation of Higher Institutions*, published by University of Chicago Press

²² George F. Zook and M. E. Haggerty *Principles of Accrediting Higher Institutions*, pages 150-151. Chicago: University of Chicago Press, 1934

The facilities and activities of an institution will be judged in terms of the purposes it seeks to serve.

C. Evaluating Various Aspects of the School

The philosophy of the school. There is rather remarkable agreement among the foregoing principles for evaluating the three levels of education, but nowhere is the agreement more notable than upon the point that an institution must be appraised in terms of its own philosophy and objectives. The Cooperative Study, for example, recommends that "a secondary school be studied expressly in terms of its own philosophy of education, its individually stated purposes and objectives, the nature of the pupils with whom it has to deal, the needs of the community which it serves, and

II. Philosophy of Secondary Education

A. SIGNIFICANT POINTS OF VIEW

The material which follows is designed to secure the viewpoint of the school concerning various aspects of educational philosophy. There is no implication that any one answer is the "right" one. Preferably only one item should be checked in each group—the one with which your school is in closest agreement as a matter of fundamental belief, regardless of actual practice. Write any modification or qualification in the space provided if you feel it necessary.

Fundamental Concepts

1 The type of political organization most desirable for society is one in which—

- () a. The determination of policies is entrusted to specially trained personnel chosen by general election
- () b. Policies are determined by individuals selected by an electorate which is restricted on the basis of racial or economic status
- () c. All individuals share in the determination of policies in proportion to their abilities
- () d. All individuals have equal voice in the determination of policies
- () e. Individuals are completely subordinated to authority, and policies are determined by a minority group

Qualifications

3 The social organization most desirable is one in which—

- () a. There are groups which have special social privileges because of hereditary family connections
- () b. Social position depends upon professional, religious, racial, or nationality status
- () c. All individuals have equal social status regardless of economic, cultural, or intellectual qualifications and regardless of race or nationality
- () d. All individuals of the dominant racial or nationality group have equal social position regardless of economic, cultural, or intellectual qualifications
- () e. Social position is given to any individual who has achieved special distinction in his field

Qualifications

2 The economic organization most desirable is one in which

- () a. Individuals may retain the results of production on the assumption that public welfare will be benefited by their philanthropies
- () b. No restrictions are placed upon the right of an individual to amass wealth
- () c. Individuals may obtain wealth but are restricted by requirements of conservation of natural resources
- () d. All share equally in the products of labor and industry
- () e. Private enterprise is encouraged but with restrictions assuring the conservation of natural resources and with provisions for the distribution of a considerable portion of the results of production in the interests of the workers and of the general public

Qualifications

4 In a democracy the school should place most emphasis upon helping to prepare pupils—

- () a. To make adjustments to present social and economic conditions
- () b. To participate in the reconstruction of society
- () c. To make adjustments to meet changing conditions

Qualifications

5 In a democracy free secondary education should be provided for—

- () a. All adolescents who are not mentally or physically defective to such an extent that they cannot be educated with normal children
- () b. Only those adolescents of superior intellectual ability
- () c. Those adolescents who can profit by a college preparatory cultural disciplinary program
- () d. Only those adolescents of superior social or economic status
- () e. All adolescents

Qualifications

Figure 46. A Suggested Technique for Evaluating the Philosophy of a Secondary School (From *Evaluation Criteria*, 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 8)

the nature of the American democracy of which it is a part."²² It recognizes four distinct phases in the satisfactory evaluation of a secondary school²³

- 1 Statement by the school of its philosophy of secondary education and of its objectives
- 2 Checking and validation of the statements of philosophy and objectives against the needs of the pupil population and community which the school serves
- 3 Revision or modification of the statements of philosophy and objectives if necessary, in light of step number 2 above
- 4 Evaluation of all aspects of the school in terms of these revised statements of philosophy and objectives. This phase involves the use of the rest of the *Evaluative Criteria*

Figure 16 illustrates one of the procedures suggested for formulating the school's philosophy. Step 2 above indicates briefly the procedure to be followed in evaluating this philosophy which is largely a matter of checking it for clearness, for internal consistency, and for appropriateness to the community to be served. Regarding the pupils and the community, the basic data which are required for the external evaluation are as follows²⁵

I Basic data regarding pupils

- A Graduates and enrollment by grades and by sex
- B Number of years seniors have been in the school
- C Distribution of withdrawals according to cause
- D Age-grade distribution of pupils
- E Distribution of I Q's by grades
- F Educational intentions of seniors by sex
- G Occupational intentions of seniors by sex

II Basic data regarding the community

- A Population data for the school community
- B Occupational status of adults
- C Occupational status of youth of secondary school age
- D Educational status of adults
- E Financial resources of the school district
- F Agencies affecting education
- G Additional socio-economic information (seven items)

The educational program A satisfactory statement of the school's philosophy having been formulated a basis is now available for evaluating the educational program and organization of the school. The general point of view and procedure of the Cooperative Study is given in the "Instructions" reproduced in Figure 47.

The following list of educational "temperatures" indicates the comprehensive character of the evaluation of the educational program

²² *How to Evaluate a Secondary School* (1940 Edition) *op cit* page 63

²³ *Evaluative Criteria* (1940 Edition) page 6 Washington D. C. Cooperative Study of Secondary School Standards

²⁵ *Ibid* pages 17-28

1 Curriculum and Course of Study

General principles; curriculum development, amount of offerings, English, ancient languages, modern languages; mathematics; sciences; social studies, music; arts and crafts; industrial arts, homemaking, agriculture; business education; health and physical education for vocational shop; general evaluation.

2. Pupil Activity Program

Nature and organization, school government, home rooms, school assembly, school publications; music activities, dramatics and speech, social life, physical activities of boys, physical activities of girls, school clubs; finances, general evaluation.

Instructions

GENERAL

In checking and evaluating the various features included in this section, the underlying philosophy and expressed purposes and objectives of the school and the nature of the pupil population and community which it serves (as outlined in Sections B and C) should be kept constantly in mind. Evaluations are to be made in the light of these factors. Persons making evaluations should continually ask "Do the practices in the school being evaluated accord with the philosophy and objectives of the school and meet the needs of its pupil population and community as well as do the practices of other schools?" They should not consider the size, type, or location of the school, the financial support available, state requirements, or other local factors, except in so far as these factors may have a legitimate effect on the philosophy and objectives of the school or on the needs of the community. In later interpretation of the results of evaluations suitable allowance may be made for any of these factors but at the time of evaluation an attempt should be made to evaluate the actual program of the school regardless of necessary limitations.

The two-fold nature of the work—evaluation and stimulation to improvement—should also be kept constantly in mind. Careful discriminating judgment is essential if these purposes are to be satisfactorily served. While the attainment of a high score may be desirable it is of secondary importance. It should not be permitted to interfere with accurate evaluation, otherwise, real improvement cannot be undertaken and attained.

Those making evaluations should be constantly on guard against the common tendency to choose the higher of two possible evaluations when in doubt. Unless a superior evaluation is definitely indicated and justified by available evidence, one of average or below average should be made.

CHECKLISTS

The checklists consist of provisions, conditions or characteristics found in good secondary schools. Not all of them are necessary, or even desirable, in every good school. Nor do these lists contain all that is desirable in a good school. A school may therefore lack some of the items listed but have other compensating features.

The use of the checklists requires four symbols. (1) If the provision or provisions called for in a given item of the checklist are definitely made or if the conditions indicated are present to a very satisfactory degree, mark the item, with the parentheses preceding it, with the symbol (+). (2) If the provision is only fairly well made or the conditions are only fairly well met, mark the item with the symbol (~), (3) if the provisions or conditions are needed but are not made, or are very poorly made, or are not present to any significant degree, mark the item with the symbol (0). (4) If it is unnecessary or unwise for the school to have or to supply what specific items call for, mark such items with the symbol (N). (Note: The figures are to be regarded merely as convenient symbols, not mathematical terms.) In brief, mark item

- + condition or provision is present or made to a very satisfactory degree
- ~ condition or provision is present to some extent or only fairly well made
- 0 condition or provision is not present or is not satisfactory
- N condition or provision does not apply

Space is provided at the end of each checklist for writing in additional items.

EVALUATIONS

Evaluations are to be made, wherever called for, on the basis of personal observation and judgment, in the light of the checklist as marked in accordance with the above instructions, and of all other available evidence, using a five point rating scale, as follows. (Note: The figures are to be regarded merely as convenient symbols, not mathematical quantities.)

- 5—Very superior, the provisions or conditions are present and functioning to the extent found in approximately the best 10% of regionally-accredited schools.¹
- 4—Superior, the provisions or conditions are present and functioning to the extent found in approximately the next 20% of regionally-accredited schools.¹
- 3—Average, the provisions or conditions are present and functioning to the extent found in approximately the middle 40% of regionally-accredited schools.¹
- 2—Inferior, the provisions or conditions are present and functioning to the extent found in approximately the next 20% of regionally-accredited schools.¹
- 1—Very inferior, the provisions or conditions are present and functioning to the extent found in approximately the lowest 10% of regionally-accredited schools.¹
- N—Does not apply. (When this symbol is used, explanation as to the reason the section does not apply should be given under Comments.)

Under Comments make notations of compensating features or particular shortcomings, explain those justifications of evaluations, or other pertinent matters.

¹ The definitions are given in terms of regionally-accredited schools since the Cooperative Study's experimental program is involved primarily regionally-accredited schools. If some other basis of comparison is used, the scores developed from the experimental program will not be applicable.

Figure 47. Instructions for Using the Evaluative Criteria Developed by the Cooperative Study of Secondary School Standards (From *Evaluative Criteria* 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 30)

3 Library Service

Library staff, organization and administration, book collection, number of titles, book collection, recency, book collection, general adequacy, periodicals, supplementary materials, selection of materials, teachers and the library, use by pupils, general evaluation

4 Guidance Service

Nature and organization, guidance staff, information about pupils, guidance procedures, phases of guidance, results, general evaluation

5 Instruction

Classroom activities, use of community, textbooks, methods of appraisal, special committee judgment, general evaluation

SUMMARY OF EVALUATIVE CRITERIA

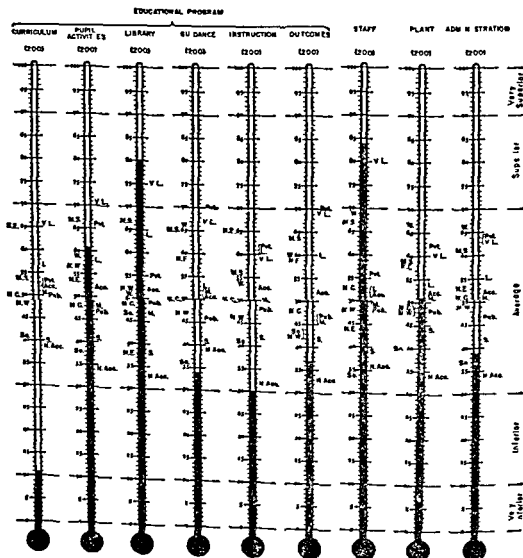


Figure 48 Summary of Evaluative Criteria for the Medium Secondary School
(From *How to Evaluate a Secondary School* 1940 Edition Cooperative Study of Secondary School Standards Washington, page 97)

6 Outcomes

Evaluation procedures; attainment in the principal subject matter fields, attitudes and appreciations

Figure 48 illustrates the graphical summary of these "temperatures" for the median school. It will be noted that this school, which happens to be a large public school, is rated "average" (between the 30th and 70th percentiles) in four of the six areas of the educational program. The school is only at the 11th percentile, or "inferior," in the curriculum, however, and at the 80th percentile, or "superior," in library. The graphical device employed makes it possible to see at a glance the strong and weak points of the program.

B. CONTENT OF OFFERINGS

In evaluating this page consider only content of subject matter offerings, not instructional procedures or methodology. Content should, however, provide not only for informational or factual matter and for skills, but also for understanding the significance of the content and for attitudes, appreciations, and ideals.

A copy of the school's courses of study should be supplied if available. If not, a brief description or outline for each course should be furnished. If the textbook serves as the course of study, it should be evaluated below.

Include in the table only those subjects or courses in which a class is taught every year or in alternate years.

If there are subjects or fields which cannot be classified in the table below, write them in the blank headings or overwrite the headings in one of the columns.

Note that the symbol "N" "condition or provision does not apply" should be used in the checklist items and evaluations of this table whenever the subject field should not be expected to contribute to the indicated item, or when the subject field is not and should not be offered in the school.

Content in each major field or area provision is made for	English	Ancient languages	Modern languages	Mathematics	Sciences	Social studies	Music	Arts and crafts	Industrial arts	Home- making	Agriculture	Business education	Physical education	Health education	Library education	Vocational shop
1. Stating the objectives to be attained																
2. Emphasizing significant contributions of our social heritage to present day life values																
3. Promoting pupils' understanding of present day social problems																
4. Stimulating pupils' interests and satisfying their needs																
5. Modifying courses to meet individual differences																
6. Including materials and experiences of potential use to adult life																
7. Interrelating the work in different subject fields																
8. Suggesting methods to be used in attaining objectives																
9. Indicating materials to be used or activities to be carried out																
10. Solving appropriate problems requiring elementary research procedures																
11. Formulating procedures for evaluating outcomes																
Evaluations																
a. How well does each course of study accord with the philosophy and objectives of the school?																
b. How well is provision made in each course of study to meet the needs of the pupil population of the school?																
c. How well is provision made in each course for correlation with other appropriate fields?																
d. How well does each course of study provide for applications to real school life?																

Comments

Figure 49. An Evaluative Procedure for the Content of the Offerings in the Principal Subject Matter Fields of a Secondary School (From *Evaluative Criteria* 1940 Edition, Cooperative Study of Secondary School Standards, page 35)

Figure 49 shows the checklist and evaluations proposed for the content of the offerings in the principal subject-matter fields. Of these, only the two with the double stars at the bottom of the columns are included in the short Gamma Scale of 25 thermometers. These two are also included in the Beta Scale, together with the three additional fields indicated by single stars. The others appear only in the complete Alpha Scale of 110 thermometers.

Bruner¹⁶ proposed an elaborate set of criteria for judging courses of study. A gross scale of four points, Excellent, Good, Fair, and Poor, is provided. The following ten questions, for example, are suggested for judging the extent to which the course of study is based upon psychological principles of learning.¹⁷

- 1 Is each new learning act considered to be in some degree remaking the whole organism?
- 2 Is self activity considered fundamental to learning?
- 3 Is study conceived of as an attack upon the situation, "and what is learned is learned as and because it is needed for the control of this situation?"
- 4 Are provisions made for taking into consideration the underlying principles of integration?
- 5 Are the activities and materials organized into patterns which, if used assist in the better growing of the individual?
- 6 Is the position held that the learner should experience satisfaction from engaging in activities?
- 7 Is knowledge considered as a means to enable the individual to participate more effectively in life situations?
- 8 Is significance attached to pupil meanings and insights?
- 9 Is the view held that growth and learning are continuous throughout the life of the individual?
- 10 Is provision made for making the situations of the school real and dramatic?

The Cooperative Study suggests a variety of procedures for evaluating the library service of the secondary school. On the assumption that the library service should be a center of the educational life of the school and not merely a collection of books, it is asserted that adequate provisions for the school library should include the following.²³

(1) a well educated efficient librarian (2) books and periodicals to supply the needs for reference, research and cultural and inspirational reading, (3) provision for keeping all materials fully cataloged and well organized, (4) a budget which provides adequately for the maintenance and improvement of the library, (5) en-

¹⁶ Herbert B. Bruner. *Criteria for Evaluating Course-of-Study Materials*. *Teachers College Record* 39: 107-120. November 1937.

¹⁷ *Ibid.* page 111.

²³ *Evaluation Criteria* (1940 Edition) *op cit*, page 51.

couragement of the pupils in the development of the habit of reading and enjoying books and periodicals of good quality and real value

Figure 50 illustrates the derivation of three measures of the adequacy of the book collection. It will be noted that books of the various classifications are weighted unequally in obtaining the composites. The two extremes are books on philosophy, with a weight of 1, and books on history, travel, and biography, with a weight of 20.

Figure 51 shows the section on Teachers and Libraries and illustrates a different technique. This section seeks answers to two important questions. First, how extensively do the teachers make personal use of the resources of the library in promoting their own professional growth and in their

III Adequacy of Library Materials

A. BOOK COLLECTION

Include books cataloged and accessioned in the library regardless of where housed. Columns F and H should not be filled out until a year the school's evaluation has been reviewed by a visiting committee. If there is to be no review, instructions for using this form will be found in *How to Evaluate a Secondary School*, pages 76-77.

Classification	Number of different titles	Number of duplicate copies	Number of titles in "Open Catalog"	Number of titles of books acquired in last year	EE EVALUATION OF New editions and duplicate editions to be counted	Average evaluation of each group	Weight to be given to each group	Weighted evaluation (product of columns F & G)	# Number of older titles (Summation from column A)	Recency Coefficient (Summation from column D)
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
000 General	3	2	1	XXX	4	3.7	10	370	8	XXX
Dictionaries	3	2	1	XXX	4					
Encyclopedias	3	0	0	XXX	3					
References	3	0	0	XXX	3	3.0	1	30	7	XXX
Other reference	3	0	0	XXX	3					
100 Philosophy	1	0	0	XXX	1					
200 Religion	15	5	5	5	4	3.6	10	360	116	59
200 Social Science	11	3	7	10	3					
Economics	43	8	15	25	4					
Pol. Sci. & Govt.	24	6	13	14	4					
Education	19	3	6	5	3					
Others	0	0	0	XXX	1					
400 Philosophy	13	2	4	9	3	3.0	15	450	64	47
500 Natural Science	21	3	6	18	3					
Physics	9	0	5	9	2					
Chemistry	12	2	3	7	3					
Biology	9	3	2	4	4					
Others	43	10	15	XXX	5					
600 Earth & Air	78	18	35	XXX	5	4.8	10	480	361	XXX
Agriculture	80	21	11	XXX	5					
Home Economics	105	26	28	XXX	4					
Business	55	10	20	XXX	5					
Others	2	0	1	XXX	1					
700 Fine Arts	1	0	0	XXX	1	1.0	5	50	3	XXX
Music	1	0	0	XXX	1					
Art	0	0	0	XXX	1					
Others	108	35	49	XXX	3					
800 Literature	23	2	19	XXX	5	3.5	15	525	182	XXX
Eng. & Amer.	30	8	6	XXX	4					
German	6	1	2	XXX	4					
French	10	4	5	XXX	3					
Spanish	4	0	1	XXX	2					
Latin	68	12	48	XXX	2					
900 History, Travel, Biography	5	0	2	XXX	1	1.0	20	200	68	XXX
Fiction	5	0	2	XXX	1					
Totals	2805	815	106							
Drivers	00									
Quantiles	2.8									
School Score										
School Score										

* Do not include any copies listed in the first column. * If a book has two copies of one title they should be counted as one. * In column A, and only duplicate copies are counted. * If a book is listed in column A, it should be counted in column B. * If a book is listed in column A, it should be counted in column C. * If a book is listed in column A, it should be counted in column D. * If a book is listed in column A, it should be counted in column E. * If a book is listed in column A, it should be counted in column F. * If a book is listed in column A, it should be counted in column G. * If a book is listed in column A, it should be counted in column H. * If a book is listed in column A, it should be counted in column I. * If a book is listed in column A, it should be counted in column J.

Figure 50 The Computation of Three Measures of the Adequacy of the Book Collections in the Library of a Secondary School (From *How to Evaluate a Secondary School* 1940 Edition Cooperative Study of Secondary School Standards Washington page 77)

classroom planning and teaching? Second, how effectively do the teachers stimulate pupils to use the library materials?

The Cooperative Study recognized five areas of guidance responsibility

V. Teachers and Libraries

A. PERSONAL USE

CHECKLIST

- | | |
|--|---|
| <p>() 1. Teachers use school and public libraries extensively to promote their own personal and professional growth</p> <p>() 2. Teachers and supervisors use the library as a stimulus to curriculum development and enrichment</p> | <p>() 3. Teachers keep the librarian informed regarding prospective classroom demand, on the library and librarian</p> <p>() 4. Teachers use the library extensively in their classroom planning and teaching</p> <p>() 5.</p> |
|--|---|

66 EVALUATIONS

- { } *x* How extensively do teachers use libraries in classroom planning?
- { } *y* How extensively do teachers use libraries for their leisure reading?

Comments

B STIMULATION OF PUPIL USE

CHECKLIST

- | | |
|--|--|
| <p>() 1. Teachers stimulate pupils to use the library, individually or in groups, to find and organize materials on selected subjects or class projects</p> <p>() 2. Teachers help pupils in the effective use of the library, largely by means of library references needed in their classroom projects</p> <p>() 3. Teachers encourage pupils to use the library for recreational and leisure reading</p> | <p>() 4. Teachers, with the help of the librarian, use the library as a means of cultivating good study and learning habits in pupils</p> <p>() 5. Teachers and classes borrow books and other library materials for use in the classroom</p> <p>() 6. Each teacher keeps a record of the voluntary reading done by the pupils in his own field</p> <p>() 7.</p> |
|--|--|

66 EVALUATION

- () *x* How effectively do teachers stimulate pupils to use library materials?

Comments

VI. Use of Libraries by Pupils

CHECKLIST

- | | |
|---|--|
| <p>() 1. Selected pupils act as assistants in the library as a means of education and exploration in library work (The time and effort of such pupils are never exploited)</p> <p>() 2. Pupils individually and in groups commonly find the library a profitable center for classroom preparation</p> <p>() 3. Pupils use libraries extensively for leisure reading and for developing other leisure interests</p> <p>() 4. Pupils help collect useful vertical file material for the library</p> | <p>() 5. Pupil activity organizations use the library extensively in the promotion of their projects</p> <p>() 6. Pupils are learning to respect public property and to help care for it</p> <p>() 7. Pupils are learning to respect the rights of others in the library and in the use of its materials</p> <p>() 8. Pupils are learning to use other libraries in the community</p> <p>() 9. Pupils use the dormitory readingroom if available</p> <p>() 10.</p> |
|---|--|

SUPPLEMENTARY DATA

- 1 Average number of school library books circulated to pupils per month
- 2 Average number of different pupils to whom school library books circulate per month
- 3 Number of high school pupils holding public library cards

EVALUATIONS

- { } *x* How extensively do pupils use library books?
- { } *y* How extensively do pupils use periodicals?
- { } *z* How extensively do pupils use supplementary materials?

Comments

Figure 51. Evaluative Techniques for the Library Service of a Secondary School. (From *Evaluative Criteria*, 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 59)

in the secondary school. These are regarded not as distinct types of guidance but rather as phases of an interrelated unitary process. These phases, together with the number of items in the checklists and the evaluations sought, are, in summary, as follows:

- A Educational Guidance 28 items
- 1 Articulation with lower schools
- How effective are procedures for articulation with lower schools?

V. Guidance Service

SUMMARY FORM

Section	Title of measure	Pages	Computation of percentage					Computation of summary score							
			Evaluation					Total	Divisor	Score	Per centile	Weights			Weighted per centile
											Alpha	Beta	Gamma		
I	Nature and organization	63	3	2	3	2.7	8	3	2.7	38	5	—	—	—	190
II	Guidance staff	64-67	2	3	3	3.2	29	3	3.2	58	20	15	—	—	1160
III	Information about pupils	67-69	4	5	4	3.9	39	10	3.9	78	10	40	60	—	780
IV	Guidance procedures	70-71	4	3	5	3.6	18	5	3.6	72	20	—	—	—	1440
V	Phases of guidance	71-74	4	3	3	3.3	23	17	3.3	62	10	45	40	—	1860
VI	Results	75	4	5	2	2.0	6	3	2.0	20	5	—	—	—	100
VII	General evaluation on	76	3	4	4	3.7	11	3	3.7	74	10	—	—	—	740
Totals									63	100	100	100	100	6270	

Summary score (Divide by 100 unless the case N = 2 the Percentile column)
Equivalent percentile (from summary conversion table)

63
58

STANDARD CONVERSION TABLE

(For priority scores based on other scores only)

Score	Percentile
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16
17	17
18	18
19	19
20	20
21	21
22	22
23	23
24	24
25	25
26	26
27	27
28	28
29	29
30	30
31	31
32	32
33	33
34	34
35	35
36	36
37	37
38	38
39	39
40	40
41	41
42	42
43	43
44	44
45	45
46	46
47	47
48	48
49	49
50	50
51	51
52	52
53	53
54	54
55	55
56	56
57	57
58	58
59	59
60	60
61	61
62	62
63	63
64	64
65	65
66	66
67	67
68	68
69	69
70	70
71	71
72	72
73	73
74	74
75	75
76	76
77	77
78	78
79	79
80	80
81	81
82	82
83	83
84	84
85	85
86	86
87	87
88	88
89	89
90	90
91	91
92	92
93	93
94	94
95	95
96	96
97	97
98	98
99	99
100	100

SUMMARY CONVERSION TABLE

(For summary scores only)

Weighted score	Alpha	Beta	Gamma
100	100	100	100
99	99	99	99
98	98	98	98
97	97	97	97
96	96	96	96
95	95	95	95
94	94	94	94
93	93	93	93
92	92	92	92
91	91	91	91
90	90	90	90
89	89	89	89
88	88	88	88
87	87	87	87
86	86	86	86
85	85	85	85
84	84	84	84
83	83	83	83
82	82	82	82
81	81	81	81
80	80	80	80
79	79	79	79
78	78	78	78
77	77	77	77
76	76	76	76
75	75	75	75
74	74	74	74
73	73	73	73
72	72	72	72
71	71	71	71
70	70	70	70
69	69	69	69
68	68	68	68
67	67	67	67
66	66	66	66
65	65	65	65
64	64	64	64
63	63	63	63
62	62	62	62
61	61	61	61
60	60	60	60
59	59	59	59
58	58	58	58
57	57	57	57
56	56	56	56
55	55	55	55
54	54	54	54
53	53	53	53
52	52	52	52
51	51	51	51
50	50	50	50
49	49	49	49
48	48	48	48
47	47	47	47
46	46	46	46
45	45	45	45
44	44	44	44
43	43	43	43
42	42	42	42
41	41	41	41
40	40	40	40
39	39	39	39
38	38	38	38
37	37	37	37
36	36	36	36
35	35	35	35
34	34	34	34
33	33	33	33
32	32	32	32
31	31	31	31
30	30	30	30
29	29	29	29
28	28	28	28
27	27	27	27
26	26	26	26
25	25	25	25
24	24	24	24
23	23	23	23
22	22	22	22
21	21	21	21
20	20	20	20
19	19	19	19
18	18	18	18
17	17	17	17
16	16	16	16
15	15	15	15
14	14	14	14
13	13	13	13
12	12	12	12
11	11	11	11
10	10	10	10
9	9	9	9
8	8	8	8
7	7	7	7
6	6	6	6
5	5	5	5
4	4	4	4
3	3	3	3
2	2	2	2
1	1	1	1

Figure 52. The Computation of the Summary Score for the Guidance Service of a Secondary School (From *How to Evaluate a Secondary School*, 1940 Edition, Cooperative Study of Secondary School Standards Washington pages 82 and 86)

- 2 Curricular and school guidance
How adequately is guidance provided in such matters as planning a sequence of studies, remedying study difficulties, etc?
- 3 Guidance concerning the post-secondary school
How adequate are provisions for assisting pupils in choices involving the post-secondary school?
- B Vocational Guidance and Placement 14 items
How adequate are provisions for assisting pupils to make wise vocational choices? How adequate are provisions for placement and follow-up service?
- C Guidance in Use of Leisure Time 6 items
How adequately are pupils assisted in making wise choices of leisure activities?
- D Social and Civic Guidance 8 items
How adequately are pupils assisted in making wise choices in matters involving social and civic relationships?
- E Personal Guidance 7 items
How adequately are pupils assisted in making wise choices in personal matters?

Figure 52 illustrates the computation of the summary score for the guidance service of the school when the Alpha Scale is used. It will be seen that the various evaluations are entered in spaces provided and then averaged. These point scores are next expressed in percentiles by the use of the standard conversion table. These percentiles are then weighted to obtain a summary score. The equivalent percentile is found from the summary conversion table at the right of the figure. The arrows indicate the sequence of events in the use of the tables. Similar conversion tables are used for the other phases of evaluation.

The quality of instruction in the school is judged by having the work of each member of the teaching staff considered from the following points of view:²⁹

- A Classroom Activities
 - 1 The teacher's plans and activities
 - 2 Cooperation between pupils and teachers
- B Use of Community and Environment
- C Textbooks and Other Instructional Materials
 - 1 Textbooks
 - 2 Other instructional materials
- D Methods of Appraisal
- E Special Committee Judgment

Figure 53 reproduces the last pages of this evaluation and illustrates the procedure.

The philosophy underlying the evaluation of the outcomes of the educational program is clearly stated in the following guiding principles:³⁰

²⁹ *Evaluation Criteria* (1940 Edition), *op. cit.* page 160

³⁰ *Ibid.* page 83

In the educational program of a good secondary school major concern should be given to attaining desirable outcomes and to the various kinds of evidence indicating that such outcomes are being realized. It may be necessary to test some outcomes by departments or in class groups. This however should not be construed as limiting the responsibilities of all phases of the educational program including the instructional activities of teachers, pupil activity program, guidance service, library, service school plant, and school administration for the attainment of desirable outcomes. There should be evidence that teachers and pupils are happily and harmoniously cooperating in the stimulation of a wholesome curiosity about themselves and their environment. Evidence should be sought to show that pupils are securing knowledge and developing worthwhile skills, attitudes, tastes, appreciations, and habits. There should be evidence that pupils are able to make desirable choices or to exercise good judgment in the selection of friends, vocations, leisure activities, goods and services, and in other important matters which confront youth today. Evaluation of such activities involves more than determining the amount of knowledge possessed, measuring the degree of skill and testing the scope of understanding important and necessary as all these are. Among others in tangible qualities such as cooperativeness, tolerance, open-mindedness, reverence, respect for law, and self-reliance are highly desirable outcomes. Evaluation of such outcomes is by no means easy, for most of them there is no standard measure and

D. METHODS OF APPRAISAL

CHECKLIST

- | | |
|---|--|
| () 1. The teacher understands the proper use, the advantages, and the limitations of various types of tests and uses them accordingly. | () 8. The teacher uses tests to stimulate and evaluate pupils' understanding and ability to make applications of knowledge. |
| () 2. The complete testing program provides for many short tests and a few relatively long ones. | () 9. The teacher uses tests to stimulate and evaluate pupils' appreciations, attitudes, and ideals. |
| () 3. Standardized achievement tests are used as well as tests of the teacher's own construction. | () 10. Pupils use tests to evaluate their own progress both in terms of educational aims and of their own purposes. |
| () 4. Tests formulated by the teacher are so planned that they are easily and economically administered, mechanically easy for pupils to take and easy to score. | () 11. Diagnostic testing is a regular part of the teaching procedure and is followed by appropriate remedial activities. |
| () 5. Testing and measuring is an integral part of the teaching and learning program rather than an activity set apart for certain days. | () 12. Other methods of appraisal such as observations of behavior, analysis of reading interests, and rating of personality traits are used. |
| () 6. The testing and measuring program emphasizes pupil progress rather than comparison. | () 13. Results of tests are made the basis for further instruction. |
| () 7. The teacher uses tests to stimulate and evaluate progress and achievement in the development of desirable habits, skills, and knowledge. | () 14. |
| | () 15. |

• EVALUATIONS

- { } x How well are methods of appraisal adapted to the purposes intended?
 { } y How well do pupils use methods of appraisal to measure their progress?
 { } z How well do teachers use methods of appraisal for determining desirable educational outcomes?

Comments:

E. SPECIAL COMMITTEE JUDGMENT

This evaluation is to be made by the visiting committee after actual classroom visitation of the teacher.

• EVALUATION

- () s How satisfactory is the instructional work carried on by this teacher?

Comments:

Figure 53. An Evaluative Technique for the Quality of Instruction in a Secondary School (Iron Evaluation Criteria, 1940 Edition, Cooperative Study of Secondary Schools Studies, Washington, page 160)

therefore evaluation of them necessarily will be largely a matter of judgment. The difficulty of the task is no reason for avoiding it, and the importance and universality of the problems involved make it imperative that attention should be directed to the attainment of such outcomes and to their proper evaluation.

Another useful instrument "designed to serve as a basis for the appraisal of individual school systems with respect to their adaptation to current educational needs" has been prepared by Mort and Cornell.²¹ It covers much the same scope as the *Cooperative Study*, but the technique is different. Specific questions are raised, to be answered *Yes* or *No*, with places for the supporting data at the left of the page. The scores for each section are then entered on a special score sheet, and by a simple process of weighting are combined into a single score. Table 43 gives the summary of this score sheet. Tentative norms are available for school systems of various sizes located in four states. The first of the ten parts of the section, which attempts to determine the degree to which the educational program recognizes the nature and extent of individual differences in pupils, is as follows:²²

- a. Intelligence Tests. Group and individual intelligence tests should be used as one of the means of analyzing problems of maladjustment.

Q. *How many of your pupils have been given intelligence tests?*

Interview Principal, Guidance director, Psychologist (if any)

Observe Individual records, Test records

Evidence

- | | | |
|--|-----|----|
| 1 Individual intelligence tests have been given in special cases | Yes | No |
| 2 A group intelligence test is given to all elementary and first year high school pupils at least once in three years | Yes | No |
| 3 Intelligence tests results for tests given both in elementary and in high school, are made a part of the permanent record of the child | Yes | No |
| 4 Educational and intelligence tests results (group and individual) constitute one of the elements upon which guidance is based | Yes | No |

²¹ Paul R. Mort and Francis G. Cornell. *A Guide for Self Appraisal of School Systems*. 59 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1937.

²² *Ibid.*, page 26.

The school organization and plant Numerous checklists and score cards have been prepared for rating school buildings equipment and administrative procedures In this field the pioneer work of George D Strayer, N L Engelhardt and their students²² is especially notable Two developments will be described briefly The scope of the evaluation procedures devised by the Cooperative Study is apparent from the following outlines of evaluative criteria

A School Staff

- 1 Numerical adequacy
- 2 Professional staff selection qualifications improvement
- 3 Nonprofessional staffs qualifications improvement in and conditions of service
- 4 Special characteristics of the school staff
- 5 General evaluation

B School Plant

- 1 The site health and safety economy and efficiency influence on the educational program
- 2 The building health and safety economy and efficiency influence on the educational program
- 3 Equipment health and safety economy and efficiency influence on the educational program
- 4 Special services cafeterias clinics etc
- 5 Special characteristics of the school plant
- 6 General evaluation

C School Administration

- 1 Administrative staff numerical adequacy preparation and qualifications improvement in service
- 2 Organization board of control general policies superintendent of schools principal
- 3 Supervision of instruction objectives procedures and activities principles results
- 4 Supervision of special services
- 5 Business management general duties and procedure budget accounting maintenance and operation
- 6 School and community relations
- 7 Special characteristics of the school administration
- 8 General evaluation

That the scope of the analysis proposed by Mort and Cornell is somewhat similar is apparent from Table 40 Short sections relating to the school site will illustrate the differences in the two techniques

²² Published by Bureau of Publications Teachers College Columbia University New York

TABLE 40

SUMMARY OF THE MOST COMPREHENSIVE SCORE SHEET FOR THE SELF APPRAISAL OF
SCHOOL SYSTEMS

Section	Adjustments Possible		Maximum Score	
	Section	Total	Section	Total
I Classroom Instruction		30		210
A The curriculum				
1 Flexibility of curriculum	10		70	
2 Breadth of curriculum	10		70	
3 Courses of study	10		70	
B Pupil activity		28		196
1 Fields of learning	13		91	
2 Extracurricular activities	7		49	
3 Instructional materials	8		56	
II Special Services for Individual Pupils		13		104
A Pupil records and attendance				
1 Educational accounting	7		56	
2 Census and attendance	6		48	
B Provisions for individual differences		26		156
1 Guidance, educational and vocational	7		42	
2 The individual and the educational program	10		60	
3 Health service	9		54	
III Educational Leadership		21		105
A Supervision and school organization				
1 Professionalization of personnel	8		40	
2 Supervision of instruction	8		40	
3 Grade and subject organization	5		25	
B School administration and the community		21		105
1 Administrative planning	6		30	
2 Status of control	7		35	
3 Scope of school influence in the community	8		40	
IV Physical Facilities and Business Management				
A The school plant		30		90
1 School plant planning	5		15	
2 The school site	5		15	
3 School buildings	10		30	
4 Special rooms	10		30	
B Business management		14		42
1 Supplies and equipment	7		21	
2 Financial accounting	7		21	
Total		183		1 008

B ECONOMY AND EFFICIENCY¹⁴

CHECKLIST

- () 1 The site is readily accessible to the school population
- () 2 It is accessible over hard surfaced roads and adequate walks
- () 3 It is sufficiently extensive for building and play needs, driveways, and landscaping
- () 4 Play areas are readily accessible
- () 5 The site has possibility of future expansion, extension or adaptation without too great cost
- () 6 It is as near the center of the school population as environmental conditions make advisable
- () 7
- () 8

EVALUATIONS

- () x How accessible is the site?
- () y How extensive is the site?
- () z How well adapted is the site for future expansion?

Comments

THE SCHOOL SITE¹⁵

- d Adaptability Each school site should be laid out and developed in consideration of both present and estimated future needs

Yes No

- Q What is the average size of the elementary school sites? Of high school sites? How do you justify the size?

Interview Superintendent
Observe Plans and data on future growth, areas of present sites and enrollment of schools

- 1 The superintendent has developed plans for the layout of each permanent site in terms of estimated changes in enrollment and educational program

Yes No

Evidence

- 2 Present development of sites is such that adjustments can be made with minimum of cost to expanding enrollment and program

Yes No

An index of variation. N. L. Engelhardt, Jr.,¹⁶ has prepared an *index of variation*, which is based upon the assumption that any true equalization of educational opportunities must provide for variability, rather than uni-

¹⁴ *Evaluative Criteria* (1940 Edition) *op cit* page 116

¹⁵ *A Guide for Self Appraisal of School Systems* *op cit* page 49

¹⁶ *The Report of A Survey of the Public Schools of Pittsburgh Pennsylvania* page 437-444 New York Bureau of Publications Teachers College Columbia University 1940 Also see *The American School and University Yearbook* for 1940

formity, in the school plant and program. The needed variation must consider the impact upon the pupil of the physical and social characteristics of the environment, as well as the personal qualities of the people to be served. For example, the thirty-eight factors that should be taken into account in determining the educational program of a community relate to the age distribution of the population, the health conditions, the housing conditions, and the social conditions of the community.

Concluding statement. Evaluation is by no means a new idea in education, although the concept has been greatly enlarged in recent years. Many new techniques have been devised to supplement those already in existence, and in some cases to supplant them altogether. Much remains to be done, however. In the meantime educators should acquaint themselves with the uses and limitations of the techniques which have been developed. There is no escaping the fact that evaluation is one of the most difficult, as well as one of the most important, problems in the modern school. The best existing evidence that a school is good is the fact that it is continually studying to find ways to improve itself.

SELECTED REFERENCES FOR FURTHER READING

- Benson, Arthur L., *How to Use the Criteria for Evaluating Guidance Programs in Secondary Schools, Form B*. Washington, D. C.: U. S. Office of Education, March, 1949.
- Cornell, Francis G., Lindvall, Carl M., and Saupe, Joe L., "An Exploratory Measurement of Schools and Classrooms," *University of Illinois Bulletin*, 50:1-71, No. 75, June, 1953.
- Dailey, John T., "Development and Application of Tests of Educational Achievement Outside the Schools," *Review of Educational Research*, 23:102-109, February, 1953.
- Domas, Simeon J., and Tiedeman, David V., "Teacher Competence: an Annotated Bibliography," *Journal of Experimental Education*, 19:101-218, December, 1950.
- Jahoda, Marie, Deutsch, Morton, and Cook, Stuart W., *Research Methods in Social Relations*. New York: Dryden Press, 1951. Chapters 5 and 14, "Data Collection: Observational Methods" and "Observational Field-Work Methods."
- Leonard, J. Paul, and Eulich, Alvin C., Editors, *An Evaluation of Modern Education*. New York: D. Appleton Century Company, 1942. 299 pages.
- Pace, C. Robert, and Browne, Arthur D., "Trend and Survey Studies," *Review of Educational Research*, 21:337-349, December, 1951.
- Reavis, W. C., and Cooper, D. H., *Evaluation of Teacher Merit in City School Systems*. Chicago: University of Chicago Press, 1945. 139 pages.
- Sells, Saul B., and Ellis, Robert W., "Observational Procedures Used in Research," *Review of Educational Research*, 21:432-449, December, 1951.
- Traxler, Arthur E. (Editor), "Measurement and Evaluation in the Improvement of Education," *American Council on Education Studies Series I*, No. 46, Vol. XV, April, 1951. 141 pages. See especially pages 58-67, "Planning a Comprehensive Evaluation Program," by Paul B. Diederich.

- Trover, Maurice E, and Pace, C Robert, *Evaluation in Teacher Education* Washington D C, American Council on Education, 1944 368 pages
- Whitney, Frederick L, *The Elements of Research* (Third Edition) New York Prentice-Hall, Inc, 1950 Appendix IV, "Representative Federal Surveys of Education "

16

Public Relations

A The Problem

The following remarks by Robert I. Thorndike, though not pertaining directly to school situations might well be applied to them.¹

In personnel selection as in most fields there is no lack of polished individuals who present in a compelling manner some completely unscientific and unvalidated technique. It is often true unfortunately that the best salesmanship is applied to the poorest product. The temperament which is disposed to careful and exacting research tends not to take kindly to or have a gift for promotion. But it is just this sound scientific worker who must in self-defense develop effective promotion for his service. The layman does not have the background to discriminate between effective personnel research and quackery. He must be educated and trained to discriminate between the tested results of a sound personnel system and the unfounded claims of the quack. The more scientific and rigorous a personnel research worker is the more important it is for him carefully to consider the public relations side of his work.

Thirty six years earlier his father and Kandel had quoted a nineteenth century educator writing in a similar vein:

Much of the scepticism prevalent as to the power and value of popular education arises from the inability of the educationist or of the school teacher to adduce satisfactory statistical evidence of the moral or of the intellectual results from any special courses of instruction or training as manifested in after life.²

¹ Robert L. Thorndike *Personnel Selection Test and Measurement Techniques* page 313 New York John Wiley & Sons 1949

² Edward I. Thorndike and Isaac L. Kandel in *Educational Measurement of Fifty Years Ago* *Journal of Educational Psychology* 4: 551 November 1913 from E. Chadwick's article in the *Modern Quarterly Magazine of Education Literature and Science* 3: 480-484 1884

Seventy five years after 1864 a survey of parent opinion of public education in America revealed deep concern over "teacher made courses of study" in which the parents have been allowed no voice, as well as concern over "the stress schools place upon the spectacular in education to the detriment of programs that will improve the manners and morals of children." These statements indicate that the public and the school do not always understand each other, and consequently do not work together to their mutual advantage.

The meaning of public relations programs. Narrowly conceived the public relations program of the school is synonymous with the publicity activities of the school. In recent years however the terms *publicity* and *propaganda* have become so closely associated and so discredited in the public mind as to arouse suspicion that something sinister is about to be 'put over.' Broadly conceived, *public relations* is merely one important aspect of the school's program of adult education. Its primary aims are two: (1) better understanding by the public of the purposes, programs, accomplishments and needs of the school and (2) better understanding by the school of the desires and needs of the community as reflected in the educational views of the public. In other words its purpose is to effect the maximum co-operation between the community's two most important educational institutions: the home and the school. And it must be remembered always that the child is the connecting link between them.

A prominent educational leader⁴ includes among the important purposes of measurement and evaluation in the modern school the provision of "psychological security to the school staff, to the pupils and to the parents," and a "sound basis of public relations." Concerning the latter Tyler says:

No factor is so important in establishing constructive and co-operative relations with the community as an understanding on the part of the community of the effectiveness of the school. A careful and comprehensive evaluation should provide evidence that can be widely publicized and used to inform the school community about the value of the school program. Many of the criticisms of the school expressed by the taxpayers and parents can be met and turned into constructive co-operation if concrete evidence is available regarding the accomplishments of the school.

There are several reasons for thinking that the problem is becoming increasingly important and difficult as the years go by. The enlarged enrollment in the secondary school and the accompanying expansion in the school program have brought many changes which the public does not understand. This fact is mainly responsible for the common charge that the modern school curriculum is cluttered up with all sorts of useless fads.

⁴ Lester S. Evans, "What Parents Expect of the School," *Journal of the National Educational Association*, 28, 194, October 1939.

⁵ Ralph W. Tyler, "The Place of Evaluation in Modern Education," *Elementary School Journal*, 41, 19, 27, September 1940.

and frills" The increasing burden of taxation has naturally made the citizens critical of all public expenditures Since in most communities the public school system is the biggest public business it is likely to bear the brunt of the attack Nor should one overlook the stubborn fact that the enormous expansion of such enterprises as are provided for by the social security and old age pension legislation has greatly increased the competition for the taxpayer's dollar In such a situation it is especially well to keep in mind the wise statement of President Madison "A popular government without popular information or the means of acquiring it is but a prologue to a farce or a tragedy, or perhaps both"

The principal sources of "popular information" may be conveniently grouped as follows

- 1 Ordinary agencies local newspapers student publications
- 2 Official publications reports bulletins handbooks etc
- 3 Report cards and letters to parents
- 4 Miscellaneous public programs exhibits P T A etc

Each of these will now receive brief discussion

B Ordinary Agencies of Public Information

Local newspapers As a medium for bringing about desirable relations between the school and its public, the newspaper ranks high For most people it is the principal source of information, but school news as reported in the local paper is likely to be narrow in scope and lack proportion An extensive early study by Farley⁵ revealed that as a rule the patrons received least information on the school topics in which they were most interested and most information on school topics in which they were least interested Table 41 summarizes the situation⁶

It may not be surprising but it is certainly unfortunate to find that in the typical newspaper the total space devoted to the first six items in order of patrons' interest was less than half that given to extracurricular activities which stand at the bottom of the list Both the school and newspaper appear to take for granted the excellent work of the classroom, which, therefore falls in the dog bites-man category rather than in the news classification⁷ They both apparently forget that a report of the incident is the most interesting thing in the world to the owner of the dog as well as to the man who has been bitten Parents never tire of hearing good reports of their own children There is no good reason why the educational side shows should be allowed to swallow up the main tent Farley calls attention to these facts⁸

⁵ Belmont Mercer Farley *What to Tell the People about the Public Schools* 136 pages New York Bureau of Publications Teachers College Columbia University 1929

⁶ *Ibid* adapted from pages 16 and 49

⁷ For an instructive discussion of this point see Edwin J Brown *Secondary-School Administration* pages 270-271 Boston Houghton Mifflin Company 1938

⁸ *Ibid* pages 16 17

TABLE 41

THE RANK ORDERS OF THIRTEEN TOPICS OF SCHOOL NEWS ACCORDING TO THE INTERESTS OF 5 067 SCHOOL PATRONS COMPARED WITH THE SPACE DEVOTED TO THESE TOPICS BY TEN NEWSPAPERS (AFTER FARLEY)

Topics of School News	Rank According to	
	Patrons	Interests
Pupil progress and achievement	1	4
Method of instruction	2	10
Health of pupils	3	9
Courses of study	4	6
Value of education	5	12
Discipline and behavior of pupils	6	11
Teachers and school officers	7	2
Attendance	8	13
Buildings and building program	9	8
Business management and finance	10	7
Board of education and administration	11	5
Parent teacher association	12	3
Extracurricular activities	13	1

In other words, patrons wish to know *what* their children are being taught *how* they are being taught, *what results* are being achieved and how the public schools affect the physical welfare of their children. They are ready to listen to the educator tell them that the results achieved in the schools are desirable that they are achieved by efficient scientific methods that children are taught useful habits and skills, that their physical welfare is not neglected.

Student publications. Student publications should occupy a strategic position in any public relations program. They represent activities that have educational value in themselves and thus constitute important exhibits of the actual work of the school. Of these publications the school newspaper and the yearbook or annual are most important. Since they are written primarily for the pupils and patrons of the school these publications can portray the actual operation of the school program more fully than the general newspaper, which must appeal to a wider public. What the student does is always of interest to other students and to parents but examination of the student publications of most schools would probably reveal a very distorted picture of the school situation. As in the regular newspaper, the extracurricular program looms large. The reader can scarcely escape the conclusion that the school year is largely occupied with social affairs and athletics. Those who criticize public education as an exponent of "fads and frills" could hardly do better than introduce the year book as Exhibit A. Beside the stadium, the library dwindles into insignificance, and such things as classrooms and laboratories are deemed so unimportant as to be omitted altogether. It is not too much to expect that the student publications present a truer picture of the school giving greater prominence to those features which justify its existence. That the public

is genuinely interested in these, there can be no doubt. Certainly parents would put evidence of pupil progress and achievement at the top of the list.

C. Official Publications

Annual reports The earliest record of a formal written educational report was made in Boston, Massachusetts, in 1738, although informal oral reports had been made to town meetings in New England at an earlier period.⁹ It is clear that from the outset the primary function of such reports has been to inform the public regarding the aims, progress, and needs of the schools and to afford an intelligent basis for determining educational policies. The first written report, for example, gave the enrollment in each school and included comments by the visiting committee on the quality of instruction. The function of such reports was well stated in the introductory pages of the 1841-1842 report of Fall River, Massachusetts, as follows:¹⁰

Those who are taxed to support Public Schools have a right to know how their money is expended and what is the character of the schools which they are required to maintain. The committee are but the agents employed by the town to take the agency of Common School Education and the employer ought to be made acquainted with all that appertains to his interest in respect to this agency. What the committee knows as to the schools the town ought to know.

Since the appearance of standardized tests, the annual reports often describe the tests used and the purposes for which they are employed, and give summaries of the results. Some cities make effective use of graphs to show that progress in the tool subjects is regular from grade to grade, as well as profile charts to illustrate the use of standard tests in the diagnosis and guidance of individual pupils. There is no way better than test results to show the need for curriculum changes, guidance services, ungraded classes, and other provisions for individual differences. There can be little doubt that parents are interested in receiving not only an account of how the money for public education was spent, but also of what it bought in the way of an efficient educational program.

But most school reports have one fatal weakness. They are not read. The reason for this has been stated as follows: "Most official reports are dull. Their authors, though they have the most interesting material in the world, treat it perfunctorily, statistically, as lifeless stuff to be put away in mortuary files."¹¹ The problem with school reports, as G. Stanley Hall long ago pointed out in the case of moral education, is how to make virtue exciting.¹²

⁹ Ward G. Reeder, *An Introduction to Public School Relations*, pages 80-87. New York: The Macmillan Company, 1937.

¹⁰ Quoted from M. G. Neale, *School Reports as a Means of Securing Additional Support for Education in American Cities*, pages 4-5. Columbia, Mo.: Missouri Book Company, 1921.

¹¹ Editorial in *The New York Times*, January 4, 1926.

¹² For a good discussion see Ward G. Reeder, *op. cit.* pages 89-104.

Special reports and publications It must be recognized that nothing is great or small, good or bad, except by comparison. Because of this fact school surveys, which attempt to interpret the local schools in relation to those of other systems of similar size, are important. At times such studies made by impartial outside agencies are especially effective. It is even better, perhaps, to have a continuous self survey, and to report at strategic intervals various phases of the school program. The larger cities employ for this purpose bulletins or magazines modeled after the house organs of industrial organizations. Graphical comparisons of standardized test scores with national norms may be so reported.

A common criticism of the modern school program is that it has allowed the newer "fads and frills" to displace the older fundamental subjects. People long for "the good old days when people really learned something when they went to school." The most effective argument with which to meet such criticism is a comparison of the achievement of the older schools and the newer, or of the traditional school program and the more liberal program of today. Riley¹² made such a study in Springfield, Massachusetts, of the results of tests in 1906 that had first been given to children in the city sixty years earlier in 1846. The findings, briefly summarized below in terms of percentage of correct responses, were favorable to the later schools.

Subjects	Percentage Correct	
	1846	1905-1906
Arithmetic	29.4	60.2
Spelling	40.6	51.2
Geography	40.3	53.4

Fish¹⁴ made a somewhat similar study comparing the achievement of Boston children in 1928 with that of pupils in the city on the same tests in 1853, seventy-five years earlier. Again the results expressed in terms of errors made favored the later schools.

Subjects	Errors Made	
	1853	1928
Arithmetic	5.4	1.6
Grammar	6.5	3.1
Geography	4.4	4.2

¹² J. I. Riley, *The Springfield Tests*, Springfield, Mass.: The Holden Patent Book Company, 1908.

¹⁴ Louis J. Fish, *Examinations Seventy-Five Years Ago and Today*, Yonkers: World Book Company, 1930.

These studies suggest preserving the results of standardized tests so that at intervals of perhaps ten or twenty years they can be compared with current results on these tests, which will afford convincing evidence of trends in efficiency. A study of this type, covering achievement in Philadelphia high schools for a ten year period, has been made by Boyer and Gordon,¹⁵ and a study of arithmetic for a twelve year period in St. Louis has been made by Boss.¹⁶

D Report Cards and Letters to Parents

Trends in report cards For many years report cards have furnished the most direct line of communication between the home and the school. They have ordinarily consisted of a record of the pupil's attendance and academic achievement, expressed in teachers' marks, sent to the parent at intervals of a month or six weeks. In recent years, however, certain important changes have taken place. In a comprehensive survey of the literature relating to report cards, Messenger and Watts¹⁷ noted the following trends:

1. There is general dissatisfaction with any scheme of grading that encourages the comparison of pupils with each other.

2. If any grades are used, a scale with fewer points is favored, a three-point scale being most often recommended.

3. There is a wide-spread feeling that the schools should evaluate traits other than mere subject-matter achievement.

4. There is a clear tendency to use descriptive rather than quantitative reports.

a. Report cards are being displaced by notes or letters to parents.

6. Cards, notes, or letters are being sent at less frequent intervals and in some schools only when there is specific occasion for such communications.

7. Attempts are being made to give more detailed diagnosis of pupils' achievements.

8. Parents are being asked to cooperate in building report forms.

9. Pupils are cooperating both in devising report cards and in evaluating their own accomplishment.

A study¹⁸ of trends in nine western states indicated that these changes were more marked in the elementary than in the secondary school. This study notes a wholesome effect on the personalities of the pupils, the effect being especially marked for those of lower ability. The most noticeable effect, however, appears to be in improved teacher-pupil relationships.

There is also evidence that these newer systems of reporting are often approved by the parents. After six years' experience, for example, one writer makes this positive statement: "The letter fosters a much more co-

¹⁵ Philip A. Boyer and Hans C. Gordon, "Have High Schools Neglected Academic Achievement?" *School and Society* 49: 810-812, June 24, 1939.

¹⁶ Mabel E. Boss, "Arithmetic Then and Now," *School and Society* 51: 391-397, March 23, 1940.

¹⁷ Helen R. Messenger and Winifred Watts, "Summaries of Selected Articles on School Report Cards," *Educational Administration and Supervision* 21: 339-350, October 1936.

¹⁸ Henry H. Hartley, "Report Card Trends in West," *Nation's Schools* 24: 51-53, November 1939.

operative relation between home and school."¹⁹ Morrisett²⁰ reports a study in which the principal of a large junior high school submitted a list of forty items to the parents with the instruction to check "items in which you are most interested, that is, those items about which you would like to know more." The item "What parents can do to promote pupil accomplishment" ranked first. Other items high in the list clearly indicated that parents desired more information regarding educational and vocational guidance. The weaknesses of the older report card was just here. The information supplied to parents, even if its accuracy could be assumed, was of such a general character as to be of little help in either diagnosis or guidance, in which full co-operation with the home is most needed.

Evans²¹ has traced the evolution of the report card. He notes a definite trend away from the standardized printed card and toward a more flexible, informal report that is better adapted to local conditions and needs. There is an increasingly clear recognition that the function of reporting is *interpretation* rather than presentation, with the emphasis on *progress* rather than on status.

Hill's study of report cards. Hill²² analyzed 443 report cards from towns and cities of all sizes, representing all educational levels and practically every state. He concluded that a satisfactory report card should

- 1 Represent the true spirit, purposes, and functions of the school
- 2 Reflect educational objectives arrived at only after careful consideration and mature judgment
- 3 Change in accord with changes in educational standards and educational philosophy
- 4 Present a report of achievement that is broad enough to cover all the important educational outcomes—subject achievement, character outcomes and social adjustment, health, and use of leisure
- 5 Give an adequate picture of *causes* as well as of outcomes
- 6 Reflect a complete and sympathetic understanding of the child
- 7 Afford a means of reporting flexible enough to account for the *particular* individual abilities of each child
- 8 Give an account of pupil progress understandable and instructive to both pupil and parent
- 9 Bring about closer cooperation and greater mutual understanding of *home* and school
- 10 Provide for reciprocal reporting [That is, space for suggestions *from* the parent]
- 11 Rate achievement in relation to the basic abilities and *character*

¹⁹ V. L. Beggs, "Reporting Pupil Progress without Report Cards," *Elementary School Journal* 37: 107-114, October 1936.

²⁰ L. N. Morrisett, "Interpreting the School to the Public," *Public Relations* 48: 3, April 1933.

²¹ Robert O. Evans, *Practices, Trends and Issues in Reporting*, 98 pages, New York: Bureau of Publications, Teachers College, Columbia University, 1938.

²² George E. Hill, "The Report Card in Present Practice," *Public Relations* 51: 1, December 1935.

- 12 Rate achievement by means of valid and reliable marking systems
- 13 Conform to reasonable standards of form and appearance The report should be attractive

The ordinary report card often fails to meet the fourth requirement in the above list. It tends to neglect the less tangible but important outcomes of education reflected in social and personal qualities. One advantage of the informal report card or letter to parents is that it attempts to inform parents on all phases of pupil growth. But it is the spirit of the report rather than its form which is important. Indeed a curt note from the teacher may be worse than the usual report card. Elsbree²³ cites the following letter from a teacher to the parents of a slow learner which is a good illustration of "How to Lose Friends and Influence Parents—in the Wrong Direction":

Dear Parents:

Donald has improved in nothing except spelling and that very little.

Sincerely,

Teacher

For use in the elementary school, Hill suggests the informal report to parents, reproduced with slight modifications in Figure 54. A similar form for the second half of the semester calls attention to improvements noted, and invites further parental co-operation on other points. Neither the report itself nor the letter accompanying it makes such demands upon the teacher's time as does the personal letter, which should probably be reserved for very special occasions. It is always a good idea, of course, to apply the grease *when* the squeak appears. The letter suggested to accompany the first report is as follows:

Dear (name of parent or guardian):

Now that the semester is one-half over we wish to call your attention to your boy's school progress. The enclosed report covers four kinds of progress—progress in school subjects, health and physical condition, attendance, and school citizenship. If you would like to talk over the report, or to get more complete information on your boy's success in school, we should be glad to have you come to see us. If you can telephone us or send a note ahead of time, it will make it easier to arrange a meeting.

The upper part of the report is for you to keep for future reference. *Please return only the lower part.* We are especially anxious to get any information from you that will aid us in helping your boy make a complete success of his school work. Any information or suggestions you may wish to write will be welcome.

Sincerely yours

(Signed by teacher and principal)

²³ Willard S. Elsbree, *Pupil Progress in the Elementary School*, page 76. New York: Bureau of Publications, Teachers College, Columbia University, 1943.

REPORT FOR FIRST HALF OF THE FIRST SEMESTER

NAME _____ GRADE _____ ROOM _____

PROGRESS IN SCHOOL SUBJECTS _____ is doing *very good*

work in _____

His work is *good* in _____

His work is *poor* and needs improvement in _____

His work in these subjects would probably be improved if _____

PHYSICAL CONDITION Health habits and conditions needing attention _____

ATTENDANCE Half days absent _____ Number of times tardy _____

REMARKS _____

SCHOOL CITIZENSHIP We believe that every boy should be happy in school should take part in the life of the school, should get along well with his classmates, and should develop good habits of honesty, courtesy, neatness, consideration for the rights of others, and industry

Your boy is especially strong in _____

He could improve in _____

Tear off here and return this part of the sheet

I have examined _____'s report for September and October

Signed _____

(Parent or guardian)

REMARKS OR SUGGESTIONS _____

Figure 54 A Suggested Informal Report to Parents (After Hill)

Suggestions for letters to parents The art of writing effective letters to parents will require special training and practice To assist teachers in acquiring this necessary skill, the schools of Santa Monica, California, prepared a very helpful list of suggestions ²⁴ The list in somewhat abridged form is as follows

- 1 Begin the letter with encouraging news
- 2 Close with an attitude of optimism
- 3 Solicit the parents' cooperation in solving the problems if any exist
- 4 Speak of the child's growth—social, physical and academic
 - a Social (citizenship traits)
 - (1) Desirable traits: attention, care of property, co-operation, honesty, effort, fair play, etc.
 - (2) Undesirable traits: selfishness, wastefulness, untruthfulness, dishonesty, carelessness, etc.
 - b Physical (health conditions)
Posture, weight, vitality, etc.
 - c Academic
 - (1) Interest in school and extra-school activities
 - (2) Methods of work
 - (3) Achievements (a) Growth in knowledge, appreciation, techniques (b) list subjects in which child is making progress and those in which he is not making progress, (3) relationship of his accepted standards to his capacities
- 5 Compare the child's efforts with his own previous efforts and not with those of others
- 6 Speak of his achievements in terms of his ability to do school work
- 7 Please remember that every letter is a professional diagnosis and therefore is as sacred as any diagnosis ever made by any physician

A more elaborate 21 page manual to guide teachers in the preparation of reports to parents was prepared by the Omaha, Nebraska, school system.

The Colorado experiment Although it is true that the aim of all evaluation and reporting to parents is the complete development of the child, it is often necessary to "temporize ideals with practical considerations."

The experience of the Secondary School of Colorado State College of Education is especially instructive ²⁵ Detailed analytical evaluation sheets were tried and abandoned primarily because of the excessive amount of time required to prepare them. The use of the terms *unsatisfactory*, *satisfactory* and *honors* was given up because it was felt that any attempt to evaluate pupils both in terms of their own ability and the objectives of the curriculum is sure to involve negative reactions. Evaluations of the ordinary scale type were tried and abandoned because they afford only a par-

²⁴ *Ibid* pages 83-84

²⁵ William L. Wrinkle, *The Story of a Secondary-School Experiment in Marking and Reporting*, *Educational Administration and Supervision* 23: 481-500, October 1937.

tial report. Anecdotal records were attempted and discontinued because the teachers tended to select unusual activities and experiences instead of reporting an ordinary picture of the pupil's growth and progress. Conference meetings of counselor, teacher, and parents, although successful for a time, had to be given up because of the failure of the majority of parents to respond to the school's invitation to avail themselves of these conference opportunities. The school eventually prepared lists of "statements of trait actions" which were indicative of the pupil's attainment of such general school objectives as self-direction, social adjustment, breadth of interests, personal attractiveness, care of materials and equipment, basic reading skills, and the like. These were then evaluated on a five-point scale, *H, S, N, U, O*, indicating distinctly superior, satisfactory, needs to make improvement, unsatisfactory, and no evaluation, respectively.

The experiment continued for many years. Wrinkle²⁶ summarized the program as follows:

In the thirteen years which have elapsed, new forms and new practices have been developed, tried, scrapped, and replaced by newer forms and practices. Detailed analytical reports, scale-type evaluations, the conference plan, anecdotal reports, and check-list type reports were developed and discarded because they did not do a good job of conveying information or demanded too much time.

Repeatedly it was discovered that adequacy meant detail and detail meant forms which were impractical for use in public school situations. One criterion which resulted in the scrapping of many forms and practices including those which were successful in their use in the laboratory school was *Whatever is developed must be usable in the public schools by public school teachers*.

In May, 1945, a popular referendum was held in which all high-school students participated; the general consensus was highly favorable but several changes were proposed. For example, 99 per cent of the students thought they should always be allowed to see their scores on standardized achievement tests; also 90 per cent of the students thought that the reports to parents should show how the actual achievement compared to the expected achievement.

The University of Chicago High School System of reporting. The University of Chicago High School plan illustrates a dual system of reporting. At the end of each semester the parents receive a detailed report in terms of the specific objectives of each course and whatever comments are deemed necessary. A week or so after the detailed reports are sent out and the parents have had an opportunity to study the strengths and weaknesses of the pupil, the course marks are forwarded and are usually accepted by the parent as incidental supplementary information. Figure 55 illustrates one of the detailed semester reports in social studies. The Chicago

²⁶ William L. Wrinkle, "Reporting Pupil Progress," *Educational Leadership*, 2, 293-295, April, 1945.

THE UNIVERSITY OF CHICAGO

The Laboratory School

SEMESTER REPORT, SOCIAL STUDIES III _____

 Student _____ Date _____
 Last Name First Name

Purposes	Rating	Comments (if any)
1 Acquisition of basic information		
2 Reading skills		
<i>a</i> recognizing main ideas		
<i>b</i> recognizing pertinent data		
<i>c</i> social studies vocabulary		
3 Oral Skills		
<i>a</i> presentation of ideas		
<i>b</i> organization of ideas		
<i>c</i> adequacy of content		
4 Writing Skills		
<i>a</i> organization of ideas		
<i>b</i> adequacy of content		
5 Ability to interpret social data		
6 Ability to apply principles in new situations		
7 Interest in current affairs		
8 Courtesy and cooperation in group situations		
Habits of Work		
9 Persistence in overcoming difficulties		
10 Tendency to work independently		
11 Promptness in completing work		
12 Application during study		
13 Attention to class activities		
14 Participation in class activities		
15 Effectiveness in following directions		
Pupil's Grade _____		

Instructor _____

Figure 55 A Report Card Used at the University of Chicago High School

system may be regarded as a desirable transition between the formal report cards and the informal letter to parents

E Other Avenues of Public Information

School exhibits. There is no sounder principle of evaluation than that contained in the statement, "By their fruits ye shall know them." Exhibits afford one of the best ways of presenting the "fruits" of the school. The public is evidently interested in local, county, state, national, and international fairs and exhibitions of all types, and schools could make use of this fact. Posters and displays of pupils' work, as well as public programs of a dramatic, literary, or musical character, afford concrete demonstrations of the school's educational program. Commencement programs in which the pupils themselves play the leading roles afford an excellent opportunity for the public to see the end products of the school. In the final analysis, however, the ordinary everyday behavior of the pupil is the best evidence of the worth of the school. What the pupil *thinks* and what the pupil *says* are both important, but what the pupil *is* speaks a still more eloquent language.

School visitation. Vicarious knowledge is important, but it is usually a poor substitute for first hand experience. Whenever possible therefore the public should have an opportunity to see their school in actual operation. The school should cultivate a reputation for friendliness. The announced policy of the school should be "The latch string is always out." It is a rare parent indeed who would not rather see his own child "perform" than witness world famous actors on television. Furthermore to observe the process of upholstering a chair or fashioning a dress is inherently more interesting than merely to look at the finished product.

The parent-teacher association. The modern educator recognizes more clearly than did his predecessor that education is a continuous unified process, that several agencies contribute to its accomplishment, and that of these the home and the school are most important. It is self-evident, therefore, that there should be intelligent and wholehearted co-operation between the home and the school. The local parent teacher association seeks through mutual understanding to effect this needed co-operation. At its best, the association is a modern successor to earlier visits of teachers to the pupils' homes and of the parents to the school, both of which are increasingly difficult with the growth of the school population and with the enlargement of the area served by the individual school.

From the viewpoint of the home the association affords an opportunity for parents not only to hear about the school's program and philosophy and to see the school in actual operation but also to react to what they hear and see. The modern parent like the modern child wants to be heard as well as seen. Certainly at all times he is entitled to communicate freely in an accepting atmosphere. A free interchange of feelings and ideas may be

facilitated by the use of group techniques such as role playing, sociodrama, and leaderless group discussion.²⁷

F Mobilizing Public Opinion

Sampling the opinion of parents To what extent can the judgment of parents be utilized in the evaluation and improvement of the school? Eells²⁸ attempted to use the opinions of the parents of seniors in evaluating the secondary schools attended by their sons or daughters. He employed a five-point scale ranging from "extremely satisfactory" at one end to "extremely unsatisfactory" at the other. Twelve items were included, relating to the general quality of instruction, development of good character, training in good citizenship, guidance activities, and the like. The principal of the school personally signed and mailed to the parents of seniors in his school a double postal card containing the following message:

To the Parents of Seniors

Our school has been selected as one of two hundred high schools and other secondary schools in the United States to be critically studied and evaluated in an effort to improve the standards of secondary education throughout the country. The study is not connected in any way with the Federal Government.

One part of the plan for this national study calls for a frank evaluation of the school from the standpoint of the parents. We are asking parents of our seniors to state their honest opinions concerning certain aspects of our school as judged by the development of their children during their school life here. You are urged to express your candid judgment whether it is favorable or unfavorable. You are not asked either to praise or to defend the school only to judge it. The card need not be signed and it is to be sent directly to the headquarters of the study in Washington. I shall not see it again.

I am eager to have a hundred per cent response from the parents of pupils in this school. Won't you fill the card out and mail it promptly? Within a day or two please!

The study concluded that "the parents, on the whole, showed a marked degree of discrimination" as judged by the scattering of the ratings along the scale. Only a quarter of the ratings were "exceedingly satisfactory," and more than 7 per cent were "not very satisfactory" or "exceedingly unsatisfactory." It is significant that the guidance program of the typical school was considered least satisfactory, a judgment supported by other criteria. Yet perhaps no phase of the school program is more dependent for its success upon parental co-operation than is guidance. "Regardless of whether parents are correct in their judgments, it is important to know what these judgments are." Eells points out "for in the last analysis the

²⁷ Helpful sources of information are Jean E. Grambs, *Dynamics of Psychodrama in the Teaching Situation*, *Societry* 1, 383-399, March 1948; Herbert A. Thelen, *Human Dynamics in the Classroom*, *Journal of Social Issues* 6, 30-50, No. 2, 1950; and William Clark Trow and others, *Psychology of Group Behavior*, *Journal of Educational Psychology* 41, 322-338, October 1950.

²⁸ Walter Crosby Eells, *Judgments of Parents Concerning American Secondary Schools*, *School and Society* 46, 409-416, September 25, 1937.

parents support and control the schools " Another writer²² emphasizes the point that although it is important to discover what the public *knows* about its schools it is even more important to learn what the public *feels* about its schools

Concluding statement. It is one of the fundamental beliefs of a democracy that reliance can be placed on an enlightened public opinion. It is to achieve this end that public schools are maintained. But it is erroneous to assume that the responsibility ceases when the formal period of instruction ends. In a changing world the continued enlightenment of the adult population is increasingly recognized as a major responsibility of a democratic society. No individual or group can be expected to think or to act intelligently on any thing without the necessary information. To supply this information about the schools is the objective of the public relations program. At all times the school will do well to keep in mind the words of one of America's ablest statesmen, Abraham Lincoln:

Public sentiment is everything. With public sentiment nothing can fail, without it nothing can succeed. Consequently he who molds public sentiment goes deeper than he who enacts statutes or pronounces decisions.

SELECTED REFERENCES FOR FURTHER READING

- Elsbree, Willard S., *Pupil Progress in the Elementary School*. New York: Bureau of Publications, Teachers College, Columbia University, 1943. Chapter VIII.
- Evans, Robert O., *Practices, Trends, and Issues in Reporting to Parents on the Welfare of the Child in School*. New York: Bureau of Publications, Teachers College, Columbia University, 1938. 98 pages.
- Froehlich, Clifford P., and Darley, John G., *Studying Students, Guidance Methods for Individual Analysis*. Chicago: Science Research Associates, 1952. 411 pages.
- Rothney, John W. M., and Roens, Bert A., *Guidance of American Youth, an Experimental Study*. Cambridge, Massachusetts: Harvard University Press, 1950. 269 pages.
- Scott, William O. N., *Desirable Objectives for Public Schools—An Opinion Analysis*. Unpublished Ph.D. Dissertation, George Peabody College for Teachers, Nashville, Tennessee, 1951. 236 pages.
- Smith, Eugene R., Tyler, Ralph W., and Staff, *Appraising and Recording Student Progress*. New York: Harper & Brothers, 1942. Chapters IX–XI.
- Sykes, Gresham M., "The PTA and Parent-Teacher Conflict." *Harvard Educational Review*, 23: 86–92, Spring 1953.
- Thorndike, Robert L., *Personnel Selection, Test and Measurement Techniques*. New York: John Wiley & Sons, 1949. Chapter 11, "The Personnel Selection Program and the Public."
- Traxler, Arthur E., *Techniques of Guidance*. New York: Harper & Brothers, 1945. Chapter XIII.
- Wrinkle, William L., and Gilchrist, Robert S., *Secondary Education for American Democracy*. New York: Farrar and Rhinehart, 1942. Chapter 39.

²² Warren C. Seyfert, "What the Public Thinks of Its Schools." *School Review* 48: 417–427, June 1940.

17

Some Present Trends

In the preceding sixteen chapters and in the six appendices on pages 429-465 a multitude of measurement problems receive attention. It seems desirable, nevertheless, in this final chapter to present a brief overview of current trends.

Reliability. The extreme emphasis upon high reliability coefficients which characterized educational and psychological measurement during the 1920's and 1930's has died down, though when a decision concerning an individual's future status in a certain trait is being based upon a single test, considerable stability is needed. For predicting a criterion, several well-constructed but short and hence only moderately reliable tests are usually better than one relatively more reliable instrument. The short tests should correlate with each other as near zero as possible, but each should correlate well with the criterion to be predicted.

'Spuriously' high single-form reliability coefficients may be obtained by two different methods, administering the test to an extremely heterogeneous group or applying a split-half or Kuder-Richardson computational procedures to a highly speeded test. If the examinees upon whom any reliability coefficient is based have more variable scores (a higher standard deviation) than your testees, then the reliability coefficient secured for your group will in all likelihood be lower than theirs.

Validity. The American Psychological Association's Committee on Test Standards lists four types of validity.¹

¹ Lee J. Cronbach (Chairman). Technical Recommendations for Psychological Tests and Diagnostic Techniques. Preliminary Proposal. *American Psychologist* 7: 461-475. August 1952. Pages 467-468. Quoted with permission of the *American Psychologist* and the American Psychological Association.

1 *Criterion validity* denotes correlation between the test and subsequent criterion measures

2 *Concurrent or status validity* denotes correlation between the test and concurrent external criteria

3 *Content validity* refers to the ease in which the specific type of behavior called for in the test is the goal of training or some similar activity. An academic achievement test is most often examined for content validity

4 *Congruent [or construct] validity* is established when the investigator demonstrates what psychological attribute a test measures by showing correspondence between scores on the test and other indicators of the state or attribute

The Committee states that "the [test] manual should make clear what type of inference the validation study supports. No manual should report that 'this test is valid.' In the past, evidence that is not appropriately termed evidence of validity has been presented in the manual under that heading." The test user should ask himself and the test salesman "Valid for what?" For instance, does the test predict success in the first year of college reasonably well? Does it correlate substantially with current level of aspiration? Is it based upon a careful sampling of the content and operations in a given set of textbooks, course units or syllabi?³ How high does it correlate with similarly named tests?⁴ Of course, few tests are valid in all four of the above senses, but the user will want to be sure that the test has the kind and degree of validity he needs.

The criterion problem. In recent years the thing to be-predicted has been shown to be of crucial importance, since even the best possible test cannot predict an extremely faulty criterion well. The criterion may not be reliable enough, it may not be completely relevant, and it may be immediate or intermediate, when some more ultimate behavior needs to be predicted.⁵ As an example, take an attitude inventory which attempts to get at "good citizenship." No matter how carefully constructed this instrument is, scores obtained on it by the pupils in a given class will not correlate well with citizenship ratings assigned to them. Furthermore, the school is probably quite interested in the adult citizenship behavior of the former student, so unless the immediate criterion—the teacher's ratings—is highly correlated with the ultimate criterion, the status validity of the inventory may differ considerably from its predictive validity.

Ratings may usually be made more reliable by having several well in

³ *Ibid.* page 468

⁴ See Philip J. Rulon, *On the Validity of Educational Tests*, *Hartara Educational Review* 16: 290-296, October 1946, also available free as Test Service Notebook No. 3, World Book Company.

⁵ A study in which tests with quite different rationales but verbally similar categories intercorrelated as expected is Julian C. Stanley and Robert S. Waldrop, *Intercorrelations of Study of Values and Kuder Preference Record Scores*, *Educational and Psychological Measurement* 12: 707-719, Winter 1952.

⁶ For a comprehensive discussion of the criterion which is particularly applicable to education see Edward E. Cureton, *Validity*, Chapter 16 in E. F. Lindquist (Editor), *Educational Measurement*, Washington, D. C., American Council on Education, 1951.

formed persons rate each individual and then take the mean of their ratings. If the raters have not all had considerable opportunity to observe the ratees with respect to the characteristic being rated, however, this process may result in some loss of relevance.

Nearly all of the criteria used in predictive validity studies are intermediate success or failure in medical school, rather than competence as a practicing physician, passing or failing in flying school, instead of achievement in combat, grades in the teacher training curriculum, not performance on the job ten years after graduation, and score on the final training-school exam in lieu of competence as an automobile mechanic. Indeed, most "ultimate" criterial measures are hard to get, all too unreliable, and of doubtful relevance. This is illustrated rather dramatically by the numerous attempts to determine what a "competent teacher" is.⁶

Factor analysis. Since the early 1930's an increasingly large number of measurement specialists have worked both theoretically and practically with factor analysis. The well known Primary Mental Ability tests (PMA)⁷ had their origins in factor analyses performed by Louis L. Thurstone, whereby he used mathematical methods to identify a few "factors" (Verbal, Word Fluency, Number, Space, Memory, Reasoning, and Perceptual Speed) which accounted for most of the positive correlations among a large number of mental tests.

The Holzinger Crowder Uni-Factor Tests for Grades VII through XII are based upon factorial studies and contain verbal, spatial, numerical, and reasoning subtests. They first appeared in 1952.⁸

Factor analysis has also been used frequently to provide information concerning what a test battery such as the Differential Aptitude Tests,⁹ the Wechsler Intelligence Scale for Children (WISC),¹⁰ or the Revised Stanford Binet¹¹ is measuring.

Achievement vs. intelligence vs. aptitude tests. The old familiar classification of ability tests into three types, achievement, intelligence, and aptitude, has been challenged severely by correlational studies. This is especially true of the eight Differential Aptitude Tests,¹² which include all three kinds: Verbal Reasoning, Numerical Ability, Abstract Reasoning,

⁶ Simeon J. Dumas and David V. Tiedeman, "Teacher Competence: an Annotated Bibliography," *Journal of Experimental Education* 19: 101-218, December 1950.

⁷ Devised by Louis L. and Thelma Gwinn Thurstone and published by Science Research Associates.

⁸ Devised by Karl J. Holzinger and Norman A. Crowder and published by the World Book Company.

⁹ Jerome E. Doppelt, *The Organization of Mental Abilities in the Age Range 13 to 17*, Contributions to Education No. 962, New York: Bureau of Publications, Teachers College, Columbia University, 1950, 86 pages.

¹⁰ Elizabeth P. Hagen, "A Factor Analysis of the Wechsler Intelligence Scale for Children," *American Psychologist* 6: 297, July 1951, Abstract.

¹¹ Lyle V. Jones, "A Factor Analysis of the Stanford Binet at Four Age Levels," *Psychometrika* 14: 299-331, December 1949.

¹² Abbreviated DAT designed for Grades 8-12 and published by the Psychological Corporation.

Space Relations, Mechanical Reasoning, Clerical Speed and Accuracy, and Language Usage (Spelling and Sentences) Such a battery, with its tests standardized on the same individuals, has distinct advantages over single aptitude tests assembled by the tester into an *ad hoc* battery All percentile ranks can be compared directly without complicated subjective attempts to allow for quite different norm groups

Bennett, Seashore, and Wesman conclude from the high correlations of the DAT Verbal Reasoning and Numerical Ability tests with intelligence tests that "apparently [they] can serve most purposes for which a general mental ability test is usually given in addition to providing differential clues useful to the counselor Hence the use of the so-called intelligence test is apparently unnecessary where the *Differential Aptitude Tests* have already been used"¹²

Differential prediction A persistent guidance problem, still largely unsolved, is to estimate differential success in a variety of fields Will John "make" a better engineer than lawyer? According to his high school grades and intelligence test scores he would probably pass either curriculum in college Can the counselor organize all the information concerning John in a way which will enable the counselor to predict with a fair degree of confidence that success in one college field is more probable than in another? As currently attempted the solution is attained largely by rule-of-thumb, "common sense" procedures which rely heavily upon intuition and "arm chair validity" If John has high mechanical and scientific interest scores on the Kuder Preference Record and high Numerical Ability, Space Relations, and Mechanical Reasoning scores on the DAT, while he is somewhat lower on the Kuder persuasive category and the DAT Verbal Reasoning, Abstract Reasoning and Language Usage tests, very likely he will be counseled toward engineering instead of law

This is an unsatisfactory method however, since it relies too heavily upon assumed validity and subjective weighting of the various test scores to arrive at a "felt" probability of success in one field versus the other For some time statisticians have been evolving methods of profile and discriminatory analysis to make differential prediction objective Though this literature has barely touched educational measurement yet it does seem to have great importance for counselors Perhaps the most easily understood articles for the interested student to read are Tiedeman's and Rulon's¹⁴

¹² George K. Bennett, Harrell G. Seashore, and Alexander G. Wesman, *A Manual for the Differential Aptitude Tests* (Second Edition) page 71 New York: Psychological Corporation, 1952

¹⁴ David V. Tiedeman, "The Utility of the Discriminant Function in Psychological and Guidance Investigations," *Harvard Educational Review* 21: 71-80 Spring 1951 and David V. Tiedeman and Jack J. Sterberg, "Information Appropriate for Curriculum Guidance," *Harvard Educational Review* 22: 257-274 Fall 1952 A splendidly humorous approach is Phillip J. Rulon, "The Statistic and the Separable: A Fable for Personnel Psychology," *Personnel Psychology* 4: 99-114 Spring 1951

The "whole" child. Emphasis has shifted somewhat from strictly objective measurement of specific traits to co-operative evaluation and appraisal of the "whole" child. Such abstract characteristics as "respecting the rights of others," "participating democratically in group activities," and "developing habits of good citizenship" occupy the attention of teachers bent upon comprehensive evaluation. Grading each child in relation to the class norm is minimized in the "modern" school, where pupils compete with their own past records.

To a considerable extent this "wholistic" approach is congruent with developments in educational philosophy and psychology since 1925, though at times it has resulted in a flight to complete subjectivism, with consequent abandonment and ridicule of objective tests. Some teachers even take the setting up of objectives to be synonymous with evaluating these objectives. Cureton¹⁵ and Rulon¹⁶ have called attention to the extremely loose thinking involved in much current "evaluation." A quotation from the former sets forth this point of view.¹⁷

Among the abstractions which we must *at present* consider intrinsically invalid we find most of the action series that go to make up "worthy home membership," "good citizenship," "democratic attitude," "loyalty," and many of the other ultimate aims of education. On the other hand, "command of fundamental processes" does lead to essential agreements. We can fairly well specify the acts performed in appropriate situations by persons to be designated as having such 'command'; the acts performed in similar situations by persons who are to be labeled as lacking such 'command'; and the materials upon which the acts are to be performed and the bases upon which the acts are to be classified and scored as successful or unsuccessful. Those educators who insist (and rightly we believe) that other aims are at least equally important and in aggregate probably much more important would advance their cause most rapidly and effectively by setting about the task of specifying the materials, actions, situations, and scoring criteria implied by the abstract terms which denote these other aims. They will find the task difficult but in most cases possible. When they have accomplished it they will find that teachers will use the materials, set up appropriate school situations, and teach the desired acts. *In those few cases where the task turns out to be impossible the abstract aim must be admitted to be intrinsically invalid.*

Qualitative and semi-quantitative evaluation techniques. In response to the recent emphasis upon evaluating non intellectual aspects of the child's behavior there have arisen several new procedures. Anecdotal records, samples of revealing behavior recorded shortly after their occurrence by the observer and made a part of the child's cumulative record, are widely used by teachers. Various types of ratings—of one's self, by peers, by teachers, and by parents—have become popular.¹⁸ Some highly

¹⁵ Edward F. Cureton *op cit*

¹⁶ Phillip J. Rulon. On the Concepts of Growth and Ability. *Harvard Educational Review* 17: 1-9 Winter 1947

¹⁷ Edward E. Cureton *op cit* page 602. Italics added.

¹⁸ Sometimes self ratings are compared with test scores as in the study by Julian C. Stanley. Insight into One's Own Values. *Journal of Educational Psychology* 42: 399 405 November 1951

refined standardized rating scales have appeared, but the majority have been prepared locally. Check lists are used frequently, too.¹⁹

Somewhere between the tests of ability and sheer rating scales are the numerous inventories which employ forced-choice methods. The Allport-Vernon-Lindzey Study of Values, illustrated on pages 176 and 190, consists of questions to which there are no objectively "right" or "wrong" answers. The Kuder Preference Record has triad items, each containing three activities from which the examinee is to pick the one he likes most and the one he likes least, obviously, this is a 1-2-3 ranking arrangement. Neither of these inventories was prepared in a purely subjective manner, for both had to meet certain statistical criteria.

Various sociometric techniques have been devised for disclosing relationships among members of a group or between groups. Frequently these are of the "Which five children in the class would you rather sit next to?" type. By this means teachers add to their information concerning social isolates and popular individuals, thereby enabling them to individualize instruction more effectively. Other sociological innovations which may at times be of value to educators are role playing, psychodrama, and sociodrama.²⁰ Leaderless group discussions for the identification of leaders within small groups have been used in industry, but apparently not as yet in public schools.²¹

Projective techniques, which until recently were used chiefly with neurotic or psychotic adults, have in some instances been adapted to the study of relatively normal children. A projective instrument allows the subject to "project" his anxieties, fears, hopes, and frustrations in a partially unstructured situation where he is unaware that he is thus revealing these inner feelings. The stimulus for this projection may be a particular set of inkblots, as in the Rorschach test, a set of specially prepared ambiguous pictures of human beings, as in the Thematic Apperception Test (TAT), selected pictures of animals, as in the Children's Apperception Test (CAT), incomplete sentences, Make a Picture-Story Test, Draw a Person Test, doll play, and many other devices. For a discussion of the "Development and Applications of Projective Tests of Personality," including a 94-item

¹⁹ The student who wants to know more about ratings, questionnaires and check lists may consult Arvil S. Barr, Robert A. Davis and Palmer O. Johnson *Educational Research and Appraisal* Chicago: J. B. Lippincott Company, 1953. 362 pages.

²⁰ See George Sharp *Curriculum Development as Re-education of the Teacher* 132 pages New York: Bureau of Publications, Teachers College, Columbia University, 1951.

Arthur Singer reports in a follow up study on "Certain Aspects of Personality and Their Relation to Certain Group Modes and Constancy of Friendship Choices" *Journal of Educational Research*, 45: 33-42, September, 1951. Also pertinent is Herbert A. Thelen, "Human Dynamics in the Classroom," *Journal of Social Issues*, 6: 30-55, No. 2, 1950.

²¹ Mary E. Roseborough, "Experimental Studies of Small Groups," *Psychological Bulletin* 50: 275-303, July, 1953, and Bernard M. Bass, "An Analysis of the Leaderless Group Discussion," *Journal of Applied Psychology*, 33: 527-533, 1949.

bibliography for 1949-52, see Rothney and Heimann.²² A severe limitation of projective techniques for the teacher or administrator is that, not being tests in the usual sense, they require for proper administration and interpretation far more clinical training than the nonspecialist can hope to acquire. Especially in this area a little knowledge can be a mighty dangerous thing.

Novel tests and items. Some of the newer tests of "general mental ability" are mentioned by Stanley.²³ These include the multi-score Wechsler Intelligence Scale for Children (WISC),²⁴ an individual test which yields two separate IQ's, performance and verbal, the Arthur Adaptation of the Leiter International Performance Scale,²⁵ an untimed test given without verbal instructions which should be useful for testing young children with physical and linguistic handicaps, the Northwestern Intelligence Tests, developmental scales for infants 4-36 weeks of age that yield IQ's,²⁶ the Davis-Eells Games for Grades I-VI, meant to be "fair" to children from all socio-economic levels,²⁷ and Goossen's ingeniously disguised six-item intelligence test for public-opinion pollsters, which masquerades as an interview measure of knowledge of current events.²⁸

A widespread recent effort, energized by the Progressive Education Association's eight-year study,²⁹ has been to measure understanding rather than merely memorization. The *Forty Fifth Yearbook of the National Society for the Study of Education, Part I*, entitled "The Measurement of Understanding,"³⁰ represents a systematic attempt to outline principles helpful in constructing tests that tap this "higher" type of knowledge. Enough has already been done with such instruments as the PEA Interpretation of Data Test, the Cooperative English Test C ("Reading Comprehension"),³¹ the Tests of General Educational Development (GED),³² and the Watson-

²² John W. M. Rothney and Robert A. Heimann, *Review of Educational Research* 23 70-84 February 1953.

²³ Julian C. Stanley, *Development and Applications of Tests of General Mental Ability*, *Review of Educational Research* 23 11-32 February 1953.

²⁴ Devised by David Wechsler and published by the Psychological Corporation.

²⁵ Devised by Grace Arthur and published by the Psychological Service Center Press.

²⁶ Devised by Adam R. Gilliland and published by Houghton Mifflin Company. The use of IQ's for children less than four or five years of age is open to serious question however.

²⁷ Published by World Book Company. For mention of the studies underlying these tests turn back to page 278.

²⁸ Carl V. Goossen, *The Goossen Hidden Intelligence Test*, *Public Opinion Quarterly* 14 759-766 Winter 1950.

²⁹ Eugene R. Smith, Ralph W. Tyler, and staff, *Appraising and Recording Student Progress*, 550 pages, New York: Harper and Brothers, 1942.

³⁰ Chicago: University of Chicago Press, 1946, 338 pages.

³¹ Devised by Frederick B. Davis, Harold V. King, and Mary Willis, and published by the Cooperative Test Division of Educational Testing Service.

³² Prepared by the Examinations Staff of the United States Armed Forces Institute and distributed by the Cooperative Test Division of Educational Testing Service.

Glaser Critical Thinking Appraisal³³ to indicate clearly that a frightened retreat to the essay test is not the only way to measure understanding, or indeed, the most desirable one. When properly constructed, objective type items measure much more than merely recognition. For many purposes now served by essay and completion tests, objective-type tests would be more suitable if prepared in accordance with well known measurement principles. This is not to deny that as often constructed, objective items—especially true-false ones—may measure little of importance.

Information and decision-making. Cronbach³⁴ and others are exploring the implications for measurement of the recently developed theory of decision making.³⁵ According to their viewpoint, a test should enable one to make better decisions than he can make without the test—not just better than chance alone, since rarely do we make decisions by sheer chance. The educator is concerned with acquiring additional information upon which to base his many decisions. Does Mary need extra help with reading? Should Jean take chemistry? Shall I concentrate upon helping Joe adjust better to the group, or is he already doing well enough?

A crucial aspect of making an accurate decision centers around how much information one already has and how much more a certain test or tests can contribute. For instance, should the teacher administer an intelligence test to his class? Hubbard and Flesher³⁶ found the average correlation between teachers' estimates of pupils' intelligence and the pupils' intelligence-test scores to be .72. Hanna's³⁷ interview estimates of intelligence correlated .71 with scores on the ACE Psychological Examination and .66 with the Ohio State University Psychological Test, while the ACEPE and the OSUPT correlated .77. Thus if the teacher has plenty of testing time available, he can expect by using an intelligence test to gain some information concerning this aspect of his pupils that he does not already have and perhaps cannot acquire easily otherwise, but the increment may not be large. If, on the other hand, test time is limited (as it usually is), he may want to administer a test of some other significant characteristic, even though it be less reliable and less valid than the intelligence test.

³³ Devised by Goodwin Watson and Edward Maynard Glaser and published by World Book Company.

³⁴ Lee J. Cronbach, *A Consideration of Information Theory and Utility Theory as Tools for Psychometric Problems*, 65 pages. Urbana: Bureau of Research and Service, College of Education, University of Illinois, November 1953.

³⁵ Ward Edwards, *The Theory of Decision Making*, *Psychological Bulletin* 1954, 51: 380-417, July 1954, 209 references.

³⁶ Robert E. Hubbard and William R. Flesher, *Intelligent Teachers and Intelligence Tests—Do They Agree?*, *Educational Research Bulletin* 32: 113-122, 139-140, May 13, 1953.

³⁷ Joseph V. Hanna, *Estimating Intelligence by Interview*, *Educational and Psychological Measurement* 10: 420-430, Autumn 1950.

Suppose, for example, that the correlation between the teacher's estimates of his pupils' mental health and the best available criterion of mental health is only .05, while the mental health test correlates .30 with this criterion. Then, even though the test has what is usually interpreted as low validity, still it contributes to the teacher's very meager initial information concerning the mental health of his students. Therefore, he would appear to be well advised to give the mental health test in lieu of the intelligence test if both compete for the same time, particularly when important decisions having to do with the mental health of the pupils must be made. This approach stresses two considerations, the accuracy of the judgment that can be made without the test and the importance of the area tested.

Thus the benefit from a test is not a function only of the test itself, but also of the decisions to be made with it. The test is just one step toward the goal of efficient decision making. In the classroom situation decisions can be changed as further information is acquired. Viewed from this standpoint, deciding tentatively on the basis of prior evidence and a low score on a mental health test that Bill is having adjustment difficulties does not classify him irrevocably as maladjusted. With its validity coefficient of only .30, the test will yield quite a few "false negatives," persons who score low on it but are not poorly adjusted. These will be discovered by the alert teacher during further screening, when he works with all low scorers more closely than heretofore.

It is important to measure a variety of characteristics, even though somewhat inaccurately, to know your risk, and to follow through with subsequent checks. Interviews, essay tests, and projective tests are not rifles aimed at a narrow target, rather, they are sawed off shotguns spraying rather wildly but frequently hitting the mark, while at the same time nicking some innocent bystanders.

It is too early to tell how much this promising-appearing application of utility theory will contribute to measurement. The interested reader may follow developments in the journal literature by means of subject and author indexes in *Psychological Abstracts*.

SELECTED REFERENCES FOR FURTHER READING

- Coombs, Clyde H., *A Theory of Psychological Scaling*. Ann Arbor: Engineering Research Bulletin No. 34, University of Michigan, May, 1952. 94 pages.
- Cureton, Edward E., "The Principal Compulsions of Factor Analysts," *Test Service Notebook No. 4*, World Book Company, (1949).
- Doppelt, Jerome E., "The Correction for Guessing," pp. 1-4 in *Test Service Bulletin No. 46*, The Psychological Corporation, January, 1954. Free.
- Jahoda, Marie Deutch, Morton, and Cook, Stuart W., *Research Methods in Social Relations, with Especial Reference to Prejudice*. New York: The Dryden Press, 1951. Parts One and Two.

- Lord, Frederic, "A Theory of Test Scores," *Psychometric Monograph No. 7*, 1952
84 pages
- "Technical Recommendations for Psychological Tests and Diagnostic Techniques,"
Psychological Bulletin, 51 1-38, March, 1954

Suppose, for example, that the correlation between the teacher's estimates of his pupils' mental health and the best available criterion of mental health is only .05, while the mental health test correlates .30 with this criterion. Then, even though the test has what is usually interpreted as low validity, still it contributes to the teacher's very meager initial information concerning the mental health of his students. Therefore, he would appear to be well advised to give the mental health test in lieu of the intelligence test if both compete for the same time, particularly when important decisions having to do with the mental health of the pupils must be made. This approach stresses two considerations, the accuracy of the judgment that can be made without the test and the importance of the area tested.

Thus the benefit from a test is not a function only of the test itself, but also of the decisions to be made with it. The test is just one step toward the goal of efficient decision making. In the classroom situation decisions can be changed as further information is acquired. Viewed from this standpoint, deciding tentatively on the basis of prior evidence and a low score on a mental health test that Bill is having adjustment difficulties does not classify him irrevocably as maladjusted. With its validity coefficient of only .30, the test will yield quite a few "false negatives," persons who score low on it but are not poorly adjusted. These will be discovered by the alert teacher during further screening, when he works with all low scorers more closely than heretofore.

It is important to measure a variety of characteristics, even though somewhat inaccurately, to know your risk, and to follow through with subsequent checks. Interviews, essay tests, and projective tests are not rifles aimed at a narrow target, rather, they are sawed-off shotguns spraying rather wildly but frequently hitting the mark, while at the same time nicking some innocent bystanders.

It is too early to tell how much this promising-appearing application of utility theory will contribute to measurement. The interested reader may follow developments in the journal literature by means of subject and author indexes in *Psychological Abstracts*.

SELECTED REFERENCES FOR FURTHER READING

- Coombs, Clyde H., *A Theory of Psychological Scaling*. Ann Arbor: Engineering Research Bulletin No. 34, University of Michigan, May, 1952. 94 pages.
- Cureton, Edward E., "The Principal Compulsions of Factor Analysts," *Test Service Notebook* No. 4, World Book Company, (1949).
- Doppelt, Jerome E., "The Correction for Guessing," pp. 1-4 in *Test Service Bulletin* No. 46, The Psychological Corporation, January, 1954. Free.
- Jahoda, Marie, Deutsch, Morton, and Cook, Stuart W., *Research Methods in Social Relations, with Especial Reference to Prejudice*. New York: The Dryden Press, 1951. Parts One and Two.

- Lord, Frederic, "A Theory of Test Scores," *Psychometric Monograph No. 7*, 1952
84 pages.
- "Technical Recommendations for Psychological Tests and Diagnostic Techniques,"
Psychological Bulletin, 51, 1-38, March, 1954.

Appendices

APPENDIX

A

Fifty Questions to Help You Learn Statistics

The following multiple-choice questions are designed to help you improve your knowledge of the material in Chapter 3. Each has five options, only *one* of which is meant to be correct. Consult your book freely while answering them. Rather than write in your book, copy the question numbers on a separate sheet of paper and put after each number the letter preceding the correct option (A, B, C, D, or E).

Unless you work on the questions diligently, they will probably not increase your understanding very much. After completing them *all* as well as you possibly can, turn to pages 459-463, where answers and explanations appear. Your percentage score, corrected for chance, equals twice the number right minus one-half of the number wrong.

$100 \left(\frac{R - \frac{W}{4}}{50} \right) = 2R - \frac{1}{2}W$ An explanation of the $R - \frac{W}{4}$ formula appears on pages 156-158.

Test Scores ¹	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd</i> ²
310-319	1	5	5	25
300-309	2	4	8	32
290-299	4	3	12	36
280-289	1	2	2	4
270-279	6	1	6	6
260-269	12	0		
250-259	11	-1	-11	11
240-249	8	-2	-16	32
230-239	2	-3	-6	18
220-229	0	-4		
210-219	3	-5	-15	75
<i>N</i> = 50			$\Sigma fd = -15$	$\Sigma fd^2 = 239$

¹ This frequency distribution, illustrating the 'Calculation of SD by Use of a Guessed Average,' appears as Table 5 on page 22 of Quinn McNemar's *Psychological Statistics* (New York: John Wiley & Sons, Inc., 1949).

The First 14 of the Following Questions Refer to the Above Test Scores

- 1 If Σ means "the sum of," then $\Sigma f =$
 - A -15
 - B 11
 - C 50
 - D 239
 - E 315
- 2 The size of the interval of each class in the above distribution is
 - A 4.5
 - B 5.0
 - C 9.0
 - D 10.0
 - E 10.5
- 3 The fractional class limits of the highest class are
 - A 309.50-319.50
 - B 309.50-318.50
 - C 309.95-319.95
 - D 310.50-318.50
 - E 310.50-319.50
- 4 The midpoint of the middle class (260-269) is
 - A 259.5
 - B 264.5
 - C 265.0
 - D 269.0
 - E 269.5
- 5 The arithmetic mean,

$$\bar{M}' + \frac{1 \times \Sigma fd}{N},$$
 is
 - A 260.3
 - B 261.5
 - C 263.0
 - D 264.5
 - E 267.5
- 6 The median (50th percentile) is
 - A 258.7
 - B 259.5
 - C 260.3
 - D 264.5
 - E 267.3
- 7 The mode is
 - A 12.0
 - B 259.5
- 8 The 25th percentile (Q_1) is
 - A 230.1
 - B 240.4
 - C 244.5
 - D 248.9
 - E 249.5
- 9 The 75th percentile (Q_3) is
 - A 267.0
 - B 269.8
 - C 270.2
 - D 272.0
 - E 274.5
- 10 Q , the semi interquartile range,

$$= \frac{Q_3 - Q_1}{2} =$$
 - A 8
 - B 12
 - C 17
 - D 20
 - E 23
- 11 The 10th percentile is
 - A 245.8
 - B 240.0
 - C 239.5
 - D 239.0
 - E 234.5
- 12 The 90th percentile is
 - A 299.5
 - B 299.0
 - C 294.5
 - D 277.8
 - E 276.1
- 13 $0.4D = 0.4 \times$ (90th percentile - 10th percentile) =
 - A 20
 - B 22
 - C 25
 - D 31
 - E 55

14 Standard deviation =

$$\frac{1}{N} \times \frac{\sqrt{N \sum fd^2 - (\sum fd)^2}}{N} =$$

- A 2
- B 15
- C 17
- D 20
- E 22

15 In a frequency distribution the size of the interval of the class whose lower and upper real limits are 9.5 and 19.5 is

- A 11.0
- B 10.0
- C 9.0
- D 5.0
- E 4.5

16 In a frequency distribution the midpoint of the class whose lower and upper real limits are 99.5 and 109.5 is

- A 107.0
- B 105.0
- C 104.5
- D 102.5
- E 102.0

17 The main reason for grouping data in class intervals as a step toward carrying out calculations of statistical measures by hand (that is without using a mechanical calculator) is to

- A reduce the amount of labor involved
- B reduce the frequency of clerical errors
- C permit the calculation of measures other than the mean
- D bring out important trends in the data
- E hide the identity of the persons tested

18 In making a frequency distribution from raw data for computational purposes the first step is to

- A determine the range of scores
- B determine the whole-number limits of the classes

- C determine the real limits of the classes
- D decide upon the number of classes
- E select the class interval

19 What is the most serious criticism to be made of the following frequency distribution of test scores where the real class limits are fractional?

Whole Number Class Limits	Frequency
44-48	1
40-44	2
36-40	0
32-36	0
28-32	2
24-28	6
20-24	5
16-20	23
12-16	24
8-12	37
4-8	33
0-4	25
-4-0	3
<hr/>	
N = 161	

- A Negative scores occur
- B Thirteen classes are used
- C There are too many low scores
- D The class midpoints are not divisible by the range of scores in each interval
- E The whole-number class limits overlap

20 The 60th percentile is the point in a distribution

- A where a student has answered 40 per cent of the questions in correctly
- B which marks the distance from the median that includes 60 per cent of the cases
- C below which are 40 per cent of the cases
- D below which are 60 per cent of the cases
- E above which are 60 per cent of the cases

- 21 The midscore of the following scores (4, 6, 7, 5, 4) is

A 6.0
B 5.5
C 5.2
D 5.0
E 4.8

- 22 Given scores 46, 65, 11, 46, 46, 63 and 68. How many points difference is there between the mid score and the median when the median is computed from a frequency distribution of these scores where the class interval is 1?

A 17.00
B 0.17
C 0.63
D 0.17
E 0.00

- 23 The arithmetic mean of the following scores (4, 5, 7, 6, 4) is

A 6.0
B 5.5
C 5.2
D 5.0
E 4.8

- 24 The measure of central tendency to use when reporting data concerning wages in order to avoid the undue influence of a few extreme salaries is the

A standard deviation
B quartile deviation
C median
D range
E mean

- 25 The term "average" as used in arithmetic textbooks refers to

A variability
B the mode
C central tendency
D the median
E the arithmetic mean

- 26 From the standpoint of statistics, the term that means the same thing as "average" is

A normal
B median

C mode
D central tendency
E mean

- 27 What is the arithmetic mean of the following distribution?

Score	<i>f</i>	<i>d</i>	<i>fd</i>
0.3	7	0	
4.7	3	+1	
<hr/>			
$N = 10$			

A 3.5
B 3.2
C 2.7
D 2.4
E 2.2

- 28 $D = (P_{90} - P_{10})$. This is a measure of

A variability
B correlation
C central tendency
D averageness
E modality

- 29 The percentage of scores lying between Q_3 and the median is

A 25
B 34
C 50
D 68
E a variable quantity that depends upon the score distribution

- 30 What is the standard deviation of the following distribution?

Score	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd^2</i>
4	3			
2	4			
0	3			

A 4.00
B 4.00
C 2.40
D 2.00
E 1.55

- 31 For the following distribution, the quartile deviation or semi-inter-quartile range, Q , is

Score	f
8	1
6	2
5	4
3	4
2	3
0	2

$$N = 16$$

- A. 2.17
B. 2.08
C. 1.79
D. 1.54
E. 1.04

- 32 For the distribution in the preceding item, the median is

- A. 1.75
B. 3.00
C. 3.25
D. 3.62
E. 4.00

- 33 On a test with a standard deviation of 20 and an arithmetic mean of 80, an individual with a raw score of 70 will have a z -score of

- A. -10.0
B. -0.5
C. -0.1
D. 0.5
E. 5.0

- 34 What rank should be assigned to a score of 95 in the following distribution, if the rank of the lowest score, 93, is 9?

Score
97
97
96
95
95
95
94
94
93

- A. 5.5
B. 6.0

- C. 7.0
D. 4.0
E. 5.0

- 35 How does the mean of N consecutive untied ranks compare with the mean of N ranks in which one or more ties occur?

- A. Former is larger
B. No difference
C. Latter is larger
D. Depends upon the number of ties
E. Depends upon where the ties occur

- 36 The Pearson product moment coefficient of correlation, r_{xy} or r_{12} may vary between

- A. -2.00 and +2.00
B. -1.00 and +1.00
C. -0.92 and +0.92
D. 0.00 and +1.00
L. 0.00 and infinity

- 37 A teacher computed a correlation coefficient between scores on a reading test and scores on the *Cooperative Test of Contemporary Affairs* obtaining a value of .92. She was justified in concluding that, as measured by these two tests,

- A. knowledge of current affairs and reading ability are closely related
B. knowledge of current affairs and reading ability are unrelated to each other
C. knowledge of current affairs and reading ability are perfectly related
D. the coefficient must have been computed incorrectly
E. wide knowledge of current affairs is the result of good reading ability

- 38 Which one of these r 's has the least predictive value?

- A. .91
B. .50
C. .17
D. -.23
E. -1.00

- 39 A student computed a Pearson product-moment coefficient of correlation, r_{xy} , between paired scores in two distributions, X and Y , and found it to be 1.05. We are absolutely certain that
- A he has freakish data
 - B he should have computed Spearman's rank-difference coefficient of correlation, ρ , instead
 - C the means of the two distributions differ
 - D the correlation between X and Y is high
 - E the r has been computed incorrectly
- 40 If the X distribution is divided into 12 classes and the Y distribution is also divided into 12 classes, the number of tally marks in the scatter diagram will be
- A N
 - B $2N$
 - C 12
 - D 24
 - E 144
- 41 Arithmetic mean is to central tendency as standard deviation is to
- A average
 - B variability
 - C Q
 - D D
 - E relationship
- 42 Q_2 is to 75th percentile as median is to
- A 90th percentile
 - B 75th percentile
 - C 50th percentile
 - D 10th percentile
 - E 10th percentile
- 43 Arithmetic mean is to σ as median is to
- A Q_4
 - B Q_3
 - C Q_2
 - D Q_1
 - E Q
- 44 Frequency distribution is to median as ungrouped measures arranged in order of magnitude are to
- A range
 - B mode
 - C class interval
 - D mean
 - E midscore
- 45 $Q_3 - Q_1$ is to 50% as, for a "normal" distribution, $\text{Mean} \pm 1SD$ is to
- A 32%
 - B 34%
 - C 50%
 - D 68%
 - E 84%

Multiple-Choice Analogies (41-50)

Directions Each of the following ten items represents an analogy. In every case the first two terms of the item are related to each other in some way and the third term is related in the same way to one of the last five.

Example Shoe is to foot as hat is to

- A arm
- B hair
- C hand
- D head
- E leg

Option D, "head," is of course correct.

- 46 Arithmetic mean is to mode as σ is to
- A range
 - B median
 - C midscore
 - D D
 - E Q
- 47 Median is to point as standard deviation is to
- A volume
 - B distance
 - C square
 - D score
 - E area

48. Positive correlation is to direct as
negative correlation is to
- A. incomplete.
 - B. inconsequential.
 - C. incorrect.
 - D. inadequate.
 - E. inverse.
49. Spearman is to ρ as Pearson is to
- A. r
 - B. σ
 - C. Σ
 - D. M
 - E. Mdn
50. Rank is to order as score is to
- A. disorder.
 - B. magnitude
 - C. median
 - D. rank
 - E. variability.

APPENDIX

B¹

A Simplified Item-Analysis Procedure

A. Preparing the Items

The two characteristics usually determined for a test item are *difficulty* and *discrimination*. How hard is the item for the group tested, and how well does it distinguish between the more able and the less able students? These two aspects of an item are nearly independent of each other, the exception being that a very easy or very hard item cannot discriminate well. If all testees mark the item correctly, it has not separated the testees into two groups, the passers and the failers. Likewise, if all mark it incorrectly (or if only a chance proportion of examinees mark it correctly), the item is non-discriminating for the group.

In the following paragraphs, a simple method for analyzing items is presented and illustrated in considerable detail. Preferably, the test or subtest should contain items of only one type (for example, four-option multiple-choice). There should be a considerable number of such items—say, arbitrarily, 50 or more—and they should have been administered to a substantial number of persons. Thus the procedure works best with final examinations prepared cooperatively by several teachers and given to large groups and less well with daily or weekly tests in a single school class. Even for the latter situation it has some value, however.

Each testee should be strongly encouraged to answer *every* item for which he has any information whatsoever, even a vague hunch concerning a single one of the options.² Also, plenty of time should be available for every examinee to try every item.

The items should be arranged as nearly as possible in ascending order of difficulty. This can be done fairly well on a subjective basis, or if the item has been administered previously to a similar group, the original difficulty values may be used. Subjective estimates of relative difficulty based upon the average of independent rankings by three qualified persons usually approximate the actual ranks better than an ordering made by only one person.

¹ Before using this appendix, the student will want to read the material on pages 117-119 carefully.

² See the discussion on pages 153-154.

Test items should be prepared according to content specifications agreed upon by the teachers who will use them. If possible, each item should be typed on a 5 x 8 card, with the answer *not* indicated on the front of the card. In fact, for purposes of criticism and editing it may be well not to have the answer on the back of the card, either. A separate answer key may be better.

Each item should be constructed with great care, special attention being given to the incorrect options (called distracters, decoys, or foils). It should then be keyed and criticized on a separate sheet *independently* by each person helping to devise the test. This editing is extremely important and should be followed by a detailed conference to reconcile differences, remove ambiguities, and discard items that cannot be revised properly. Though time-consuming, a cooperative approach to test construction pays off by increasing reliability and validity and improving the morale of test-takers.

Statistical item analysis is no substitute for meticulous care in planning, constructing, criticizing, and editing items. It does supplement that intuitive process, however, by revealing unsuspected defects or virtues of the specific items.

B. A Measure of Discrimination

After the test has been given, score the papers or answer sheets by marking with a red pencil all items incorrectly answered or omitted. Because of the instructions concerning omissions they should be few. Each pupil's score will be the sum of his red marks—the smaller the better.

Divide the papers or answer sheets into three piles as follows:

1 Arrange the N papers by score, beginning on top with the smallest number (best score) and going on down to the largest number (poorest score).

2 Multiply N , the total number of testees, by 0.27 and round off the result to the nearest whole number, or look in Table 46 on pages 448–450 for the appropriate figure, called n there.

3 Count off the n best papers from the top of the stack. This is the "high" group.

4 Count off the n poorest papers from the bottom of the stack. This is the "low" group.

5 Put aside the middle group (approximately 46 per cent of the papers), since it is not used in the item-analysis.

6 Set up a form somewhat like Table 42, with Item Number, W_L , W_H , $W_L - W_H$, and $W_L + W_H$ headings.

7 W_L is the number of persons in the low group who answered a certain item wrongly, including those who omitted it. It represents the total number of red marks for that item in the low group.

8 W_H is the number of persons in the high group who answered the item wrongly including those who omitted it. It represents the total number of red marks for that item in the high group.

9 $W_L - W_H$ means " W_L minus W_H " for a given item. $W_L + W_H$ means " W_L plus W_H " for that item.

The larger $W_L - W_H$ is the more discriminating power the item has. For editing purposes it is well to arrange the items from least discriminating—and therefore in greatest need of scrutiny—to most discriminating. A few items may have negative $W_L - W_H$ values, indicating that more persons in the high than in the low group missed the item. Such items may be mis-keyed, ambiguous, or unrelated in content to the rest of the test.

For convenience a critical value of $W_L - W_H$ at or above which the item is considered suitably discriminating may be determined from Table 46 on page 448.

TABLE 42

THE 100 ITEMS IN A FIVE-OPTION MULTIPLE-CHOICE TEACHER-MADE TEST
ARRANGED ACCORDING TO DISCRIMINATING POWER, FROM THE
LEAST DISCRIMINATING TO THE MOST DISCRIMINATING

Item Number	Rank Order of Item According to Discriminating Power (1 = Poorest Discrimination)	W_L	W_H	$W_L - W_H$ (Discrimination)	$W_L + W_H$	Estimated Percentage of Examinees Who Did Not "Know" the Correct Answer to the Item* (Difficulty)
30	1	31	40	-9	71	67
35	2	1	2	-1	3	3
27	3	38	39	-1	77	73
38	4	0	0	0	0	0
31	5	38	37	1	75	71
34	6	28	26	2	54	51
42	7	4	1	3	5	5
72	8	17	14	3	31	29
32	9	5	0	5	5	5
60	10	5	0	5	5	5
29	11	9	4	5	13	12
39	12	14	9	5	23	22
94	13	6	0	6	6	6
45	14	8	1	7	9	9
28	15	49	42	7	91	86
8	16	58	51	7	109	103**
33	17	31	23	8	54	51
44	18	10	1	9	11	10
51	19	10	1	9	11	10
86	20	10	1	9	11	10
92	21	16	7	9	23	22
81	22	13	2	11	15	14
14	23	16	5	11	21	20
74	24	24	13	11	37	35
57	25	26	15	11	41	39
46	26	12	0	12	12	11
40	27	13	1	12	14	13
48	28	15	3	12	18	17
67	29	15	2	13	17	16
52	30	30	17	13	47	45
20	31	48	35	13	83	79
21	32	54	41	13	95	90
76	33	15	1	14	16	15
68	34	16	2	14	18	17
59	35	17	3	14	20	19
19	36	22	8	14	30	28
62	37	18	3	15	21	20
70	38	21	6	15	27	26
61	39	28	13	15	41	39
43	40	37	22	15	59	56

$$* \frac{100 \times O}{2n(O - 1)} (W_L + W_H) = \frac{500}{132 \times 4} (W_L + W_H) = 0.947(W_L + W_H)$$

** Slightly fewer persons answered Item No 8 correctly than would be expected on the basis of chance alone

TABLE 42 (Continued)

<i>Item Number</i>	<i>Rank Order of Item According to Discriminating Power (1 = Poorest Discrimination)</i>	<i>W_L</i>	<i>W_H</i>	<i>W_L - W_H (Discrimi- nation)</i>	<i>W_L + W_H</i>	<i>Estimated Percentage of Examinees Who Did Not "Know" the Correct Answer to the Item (Difficulty)</i>
26	43	57	42	15	99	94
49	41	16	0	16	16	15
4	15	18	2	16	20	19
82	46	18	2	16	20	19
25	47	23	7	16	30	28
77	48	18	1	17	19	18
79	49	18	1	17	19	18
75	50	19	2	17	21	20
66	51	19	1	18	20	19
58	52	23	5	18	28	27
3	53	29	11	18	40	38
18	54	25	6	19	31	29
5	55	27	8	19	35	32
23	56	39	20	19	59	56
100	57	46	27	19	73	69
36	58	51	32	19	83	79
24	59	23	3	20	26	25
37	60	25	5	20	30	28
65	61	26	6	20	32	30
22	62	45	25	20	70	66
9	63	25	4	21	29	27
56	64	35	14	21	49	46
95	65	59	38	21	97	92
7	66	53	31	22	84	80
69	67	26	3	23	29	27
47	68	31	8	23	39	37
17	69	27	14	23	51	48
91	70	26	2	24	28	27
63	71	51	27	24	78	74
1	72	30	5	25	35	33
13	73	33	8	25	41	39
16	74	33	8	25	41	39
53	75	35	10	25	45	43
10	76	38	13	25	51	48
54	77	40	15	25	55	52
64	78	27	1	26	28	27
80	79	37	11	26	48	45
55	80	42	16	26	58	55
71	81	47	21	26	68	64
84	82	30	3	27	33	31
89	83	32	5	27	37	35
98	84	36	9	27	45	43
83	85	37	10	27	47	45
11	86	42	15	27	57	54
78	87	29	1	28	30	28
97	41	37	22	15	59	56
41	42	43	28	15	71	67

TABLE 42 (Continued)

Item Number	Rank Order of Item According to Discriminating Power (1 = Poorest Discrimination)	W_L	W_H	$W_L - W_H$ (Discrimination)	$W_L + W_H$	<i>Estimated Percentage of Examinees Who Did Not "Know" the Correct Answer to the Item (Difficulty)</i>
6	88	39	11	28	50	47
96	89	36	7	29	43	41
50	90	40	11	29	51	48
85	91	52	23	29	75	71
73	92	32	2	30	34	32
99	93	43	13	30	56	53
87	94	38	7	31	45	43
93	95	54	23	31	77	73
15	96	37	5	32	42	40
88	97	37	2	35	39	37
2	98	43	8	35	51	48
12	99	41	4	37	45	43
90	100	49	8	41	57	54

Then high low group data for every option of each unsuitably discriminating item may be secured to aid in the editing process

C. A Measure of Difficulty

The larger $W_L + W_H$ is, the harder the item was for the group tested. $W_L + W_H$ may be multiplied by a constant, $\frac{100 \times O}{2n(O - 1)}$, to obtain an estimate of the difficulty of the item, corrected for chance;² here O is the number of options each item has. This approximates the percentage of the testees who did not "know" the correct answer. Items in the revised test should be arranged according to $W_L + W_H$, from lowest (easiest) to highest (hardest).

D. An Illustrative Analysis

Table 42 shows W_L , W_H , $W_L - W_H$, $W_L + W_H$, and difficulty values for each of the 100 items on an English final examination constructed by four college instructors and administered with "do-guess" instructions to 243 college freshmen at the end of the winter quarter. The item numbers have been rearranged, the least discriminating item now coming first and the most discriminating one last.

When $N = 243$, $0.27N = n = 66$, so there are 66 persons in the low group and 66 in the high group. Therefore, the maximum possible value of $W_L - W_H$ is $66 - 0 = 66$, and the minimum possible value is $0 - 66 = -66$. These figures would probably never occur except because of miskeying, however, for even by chance about $\frac{1}{3}$ th of the examinees who attempted the item would mark it correctly.

² Quite a few test experts do not favor correcting item difficulty indexes for "chance." For a discussion of this point see Frederick B. Davis, "Item Selection Techniques," Chapter 9 in E. F. Lindquist (Editor), *Educational Measurement*, pages 267-285. Washington: D. C. American Council on Education, 1951.

Thus the highest value for $W_L - W_H$ barring omissions, mis-keying, or extreme misinformation is $[66 - \frac{1}{2}(66)] - 0 = 52.8$, the lowest is -52.8 .

In practice, there will probably be few large negative values since most items have at least a little positive discriminating power. Only three negative $W_L - W_H$ figures occur in Table 42, the largest of these is -9 . The greatest positive $W_L - W_H$ value in the table is 41. It will be instructive to examine these two items (Numbers 30 and 90) carefully in order to determine why they differ radically in discriminating power. Let us take the least discriminating item first.

30 In preparing a speech the first step is to choose a subject. The speaker should then

- A. practice
- B. collect material
- C. choose gestures
- D. select main points
- E. phrase a thesis

Responses by the high and low groups (66 testees in each) were as shown in Table 43. The keyed answer was B, "collect material," but E, "phrase a thesis," appealed more to those students who earned high scores on the test as a whole. Options A and C ("practice" and "choose gestures") were practically useless, since they deceived only 3 of the 132 persons. Option D, "select main points," discriminated in the proper direction 8 to 18. The item as a whole was fairly difficult, since 67 per cent of the freshmen did not "know" the correct answer.

TABLE 43

NUMBER OF EXAMINEES IN HIGH AND LOW GROUPS WHO CHOSE EACH OPTION OF ITEM NO. 30

Group	Option						Number of Examinees
	A	B	C	D	E	Out	
High	1	26	0	8	31	0	66
Low	1	35	1	18	11	0	66
Totals	2	61	1	26	42	0	132

By using the above information, the speech teacher may be able to salvage the item without destroying its main point. He would try to determine why Option E attracted the better students. This may indicate the need for additional classroom instruction concerning steps in preparing a speech, or it may highlight a real conflict between B and E as the correct answer for the item. If the dilemma can be resolved, new distracters will then be devised to replace ineffective Options A and C.

On the other hand it may not be feasible to retain the item. Not all poorly discriminating questions can be revised successfully. Sometimes the point tested is not clear or defensible enough to serve as the basis for an item. Therefore more items for each part of the test outline should be prepared than will be needed in the revised test, so that virtually unrevisable items may be discarded. How many excess items are needed depends upon the nature of the test, the purposes for which it will be used, and the care devoted to initial construction and editing. Some items such as those concerning vocabulary, are much easier to prepare well than are others such as civics questions.

Now let us turn to the most discriminating item, No 90

"Humanity is the mould to break away from, the crust to break through, the coal to break into fire, the atom to be split" is a quotation from

- A John Dos Passos
- B Carl Sandburg
- C Robinson Jeffers
- D Kenneth Fearing
- E Sherwood Anderson

Numbers of responses to the various options are shown in Table 44, the keyed answer is C, "Robinson Jeffers." Note that all four distracters (A, B, D, E) discriminate in the right direction and reasonably well, each is more attractive to the low group. Approximately 54 per cent of the testees did not "know" the correct answer. This item does not need any editing.

How many of the items should be edited on the basis of option information like that contained in Tables 43 and 44? Probably most of them could be improved in this manner, especially by the substitution of better distracters for nonfunctioning ones, but the labor involved in this process is too great for most teachers unless only a small portion of the items are scrutinized. A rule-of-thumb procedure would be to edit the 25 per cent least discriminating items. For Table 42, where there are

TABLE 44

NUMBER OF EXAMINEES IN HIGH AND LOW GROUPS WHO CHOSE EACH OPTION OF ITEM No 90

Group	Option						Number of Examinees
	A	B	C	D	E	Omit	
High	2	1	58	3	2	0	66
Low	14	8	17	15	11	1	66
Totals	16	9	75	18	13	1	132

100 items, this involves taking the first 25 items, whose $W_L - W_H$ values are less than 12. For these 25 items information like that contained in Table 45 would be drawn up by two conscientious persons (even high school students) working together. Then the subject matter experts would edit carefully on the basis of the discrimination difficulty and option information given in Tables 42 and 45.

To provide you material upon which to practice editing, the 25 least discriminating items are presented herewith. They still contain spelling and typographical errors which appeared on the test itself. Of course, these should have been removed by careful proofing of the stencils before mimeographed copies were run off.

30 Already appears on page 441

35 The subject you choose for a talk should be

- A one that is wholly new to you
- B one that interests you but about which you know nothing
- C anything for which you can find material
- D anything which will find the required time
- E one that interests you and about which you already know something

TABLE 45

NUMBER OF EXAMINEES IN HIGH AND LOW GROUPS WHO CHOSE EACH OPTION OF THE 25 LEAST DISCRIMINATING ITEMS

Item Number	Group	Option						Number of Examinees (Check Column)
		A	B	C	D	E	Omit	
30	H	1	26	0	8	31	0	66
	L	1	35	1	18	11	0	66
35	H	1	0	1	0	64	0	66
	L	0	1	0	0	65	0	66
27	H	31	1	27	7	0	0	66
	L	26	1	28	8	3	0	66
38	H	66	0	0	0	0	0	66
	L	66	0	0	0	0	0	66
31	H	29	1	28	3	5	0	66
	L	28	2	25	5	6	0	66
34	H	8	0	40	18	0	0	66
	L	10	5	38	12	0	1	66
42	H	0	0	65	1	0	0	66
	L	0	4	62	0	0	0	66
72	H	5	0	1	52	8	0	66
	L	9	0	4	49	3	1	66
32	H	66	0	0	0	0	0	66
	L	61	3	1	0	0	1	66
60	H	0	0	66	0	0	0	66
	L	3	1	61	1	0	0	66
29	H	3	1	0	62	0	0	66
	L	6	0	1	57	2	0	66
39	H	0	57	0	2	7	0	66
	L	0	52	5	2	7	0	66
94	H	66	0	0	0	0	0	66
	L	60	3	1	1	0	1	66
45	H	65	0	0	0	1	0	66
	L	58	1	2	1	4	0	66
28	H	0	2	7	24	33	0	66
	L	0	4	11	17	32	2	66
8	H	13	15	9	21	8	0	66
	L	8	8	6	34	10	0	66

TABLE 45 (Continued)

Item Number	Group	Option						Number of Examinees (Check Column)
		A	B	C	D	E	Omit	
33	H	1	13	21	0	1	0	66
	L	0	35	25	2	3	1	66
44	H	65	0	1	0	0	0	66
	L	56	1	8	1	0	0	66
51	H	0	65	1	0	0	0	66
	L	2	56	2	5	1	0	66
86	H	0	0	0	0	65	1	66
	L	1	2	2	4	56	1	66
92	H	0	0	1	5	59	1	66
	L	3	0	1	11	50	1	66
81	H	0	0	1	1	61	0	66
	L	0	0	1	12	53	0	66
14	H	1	61	1	3	0	0	66
	L	9	50	2	4	1	0	66
74	H	3	53	2	6	2	0	66
	L	6	42	6	9	2	1	66
57	H	9	0	4	51	2	0	66
	L	11	2	5	40	6	2	66

27 In a panel discussion the members of the panel

- A deliver prepared speeches
- B ask questions of the audience
- C provide informal discussion for audience
- D answers questions of the moderator
- E speak in rotation from right to left

38 The best material for a speech

- A holds interest and develops thesis
- B bores audience but develops thesis
- C pleases the speaker but annoys the audience
- D pleases audience but ignores purpose of speech
- E is unreliable but develops thesis

31 A speaker whose purpose is to instruct should begin by

- A showing why the information is needed
- B telling a funny story
- C stating his thesis
- D putting a diagram on the board
- E stating his qualifications to speak

34 The fundamental process under which is included methods of organizing a speech is

- A adjustment to the speaking situation

- B articulation
 - C choice of material
 - D symbolic formulation and expression
 - E phonation
- 42 Good posture for a speaker should be
- A rigid and stiff
 - B oratorical and pompous
 - C comfortable and natural
 - D odd and unusual
 - E lax and undisciplined
- 72 In "American Letter" MacLeish expresses
- A loyalty to a foreign land
 - B disgust with American industry
 - C disgust with American tradition
 - D loyalty to America
 - E a desire to leave America
- 32 In choosing a subject a speaker should try to find one which
- A he is or can become enthusiastic about
 - B he dislikes but which may please the audience
 - C will annoy his audience
 - D will require little preparation
 - E he has seen in a popular magazine
- 60 "Roman Stillborn" was written by
- A Oswald Spengler
 - B Aldous Huxley
 - C Robinson Jeffers
 - D Leonard S. Brown
 - E George Boas
- 29 In public discussion he makes the best contribution who
- A argues down all objections to his proposals
 - B does the most talking
 - C listens attentively but says nothing
 - D makes a creative adjustment between conflicting points of view
 - E refuses to permit the expression of conflicting points of view
- 39 The best kind of introduction will
- A make the audience laugh and feel good
 - B get favorable attention and lead into subject
 - C impress the audience with the importance of the speaker
 - D present major arguments to be developed
 - E introduce a need step thesis and main points of speech
- 94 The main idea in the selection in the text from Steinbeck's *The Grapes of Wrath* is to
- A show the conflict between those who own the land and those who care for it
 - B tell the story of a farmer who became a day laborer
 - C describe a family of poverty-stricken children
 - D explain the importance of rotation of crops
 - E urge tenants to pay their taxes
- 45 In order to avoid stage fright perhaps the best precaution is
- A to be thoroughly prepared
 - B to write your speech and read it
 - C to display a "don't care anyhow" attitude to the audience
 - D to display an overconfident overbearing attitude
 - E to avoid looking directly at your audience

TABLE 45 (Continued)

Item Number	Group	Option						Number of Examinees (Check Column)
		A	B	C	D	E	Omit	
33	H	1	13	21	0	1	0	66
	L	0	35	25	2	3	1	66
44	H	65	0	1	0	0	0	66
	L	56	1	8	1	0	0	66
51	H	0	65	1	0	0	0	66
	L	2	56	2	5	1	0	66
86	H	0	0	0	0	65	1	66
	L	1	2	2	4	56	1	66
92	H	0	0	1	5	59	1	66
	L	3	0	1	11	50	1	66
81	H	0	0	1	1	64	0	66
	L	0	0	1	12	53	0	66
14	H	1	61	1	3	0	0	66
	L	9	50	2	4	1	0	66
74	H	3	53	2	6	2	0	66
	L	6	42	6	9	2	1	66
57	H	9	0	4	51	2	0	66
	L	11	2	5	40	6	2	66

27 In a panel discussion the members of the panel

- A deliver prepared speeches
- B ask questions of the audience
- C provide informal discussion for audience
- D answers questions of the moderator
- E speak in rotation from right to left

38 The best material for a speech

- A holds interest and develops thesis
- B bores audience but develops thesis
- C pleases the speaker but annoys the audience
- D pleases audience but ignores purpose of speech
- E is unreliable but develops thesis

31 A speaker whose purpose is to instruct should begin by

- A showing why the information is needed
- B telling a funny story
- C stating his thesis
- D putting a diagram on the board
- E stating his qualifications to speak

34 The fundamental process under which is included methods of organizing a speech is

- A adjustment to the speaking situation

- B articulation
 - C choice of material
 - D symbolic formulation and expression
 - E phonation
- 42 Good posture for a speaker should be
- A rigid and stiff
 - B oratorical and pompous
 - C comfortable and natural
 - D odd and unusual
 - E lax and undisciplined
- 72 In "America in Letter" MacLeish expresses
- A loyalty to a foreign land
 - B disgust with American industry
 - C disgust with American tradition
 - D loyalty to America
 - E a desire to leave America
- 32 In choosing a subject a speaker should try to find one which
- A he is or can become enthusiastic about
 - B he dislikes but which may please the audience
 - C will annoy his audience
 - D will require little preparation
 - E he has seen in a popular magazine
- 60 "Roan Stillion" was written by
- A Oswald Spengler
 - B Aldous Huxley
 - C Robinson Jeffers
 - D Leonard S. Brown
 - E George Boas
- 29 In public discussion he makes the best contribution who
- A argues down all objections to his proposals
 - B does the most talking
 - C listens attentively but says nothing
 - D makes a creative adjustment between conflicting points of view
 - E refuses to permit the expression of conflicting points of view
- 39 The best kind of introduction will
- A make the audience laugh and feel good
 - B get favorable attention and lead into subject
 - C impress the audience with the importance of the speaker
 - D present major arguments to be developed
 - E introduce a need step thesis and main points of speech
- 94 The main idea in the selection in the text from Steinbeck's *The Grapes of Wrath* is to
- A show the conflict between those who own the land and those who care for it
 - B tell the story of a farmer who became a day laborer
 - C describe a family of poverty stricken children
 - D explain the importance of rotation of crops
 - E urge tenants to pay their taxes
- 45 In order to avoid stage fright perhaps the best precaution is
- A to be thoroughly prepared
 - B to write your speech and read it
 - C to display a 'don't care anyhow' attitude to the audience
 - D to display an overconfident overbearing attitude
 - E to avoid looking directly at your audience

28. The round table method of public discussion is suitable for groups of not more than:
- 100.
 - 50.
 - 25.
 - 15.
 - 5.
8. One of the following is a run-on sentence. That sentence is:
- Dinner was served, and we ate rapidly.
 - Some of the people are waiting, others have gone ahead.
 - This is the problem; the solution is clear.
 - He paused, adjusted his tie, and rang the bell; the maid refused to open the door.
 - Whistles, sirens, horns, and firecrackers broke the silence, and the New Year was born.
33. Barnes emphasises the Four Fundamental Processes of Speech. Of these the first is
- phonation.
 - adjustment to the speaking situation.
 - choice of material.
 - control of bodily activity.
 - projection to the audience.
44. Best advice in developing a lively sense of communication is
- be energetic, speak with enthusiasm.
 - be passive, apathetic.
 - appear reluctant to meet the assignment.
 - avoid looking directly at the audience.
 - speak with an outburst of oratorical display.
51. Steinbeck's *The Grapes of Wrath* pictures primarily:
- the shiftlessness of the average American farm worker.
 - the plight of the Western tenant farmers and migratory workers.
 - the effect of Communist propaganda on poor tenant farmers.
 - the immorality of the ignorant migratory workers.
 - the lack of religious faith among American farmers.
86. A boy whose fascination with a machine later turned to fear was found in
- Roan Stallion.
 - Tractored Off.
 - R. U. R.
 - Our Changing Characteristics.
 - Mr. Mechano.
92. The Wright Brothers' home was
- on Albermarle Sound.
 - at Fort Meyers.
 - in St. Petersburg, Florida.
 - on the coast of North Carolina.
 - on Hawthorne Street in Dayton, Ohio.
81. President Truman in his Fordham Address said the one defense against the atom bomb lies in
- bigger and better fighter planes.
 - an adequate air raid warning system.
 - aggressive warfare.
 - making a stronger U. N.
 - mastering the science of human relationships.

- 14 Which sentence is best punctuated?
 A I have no pencil, and I do not want one
 B I have no pencil, and I do not want one
 C I have no pencil and I do not want one
 D I have no pencil—and I do not want one
 E I have no pencil and I do not want one
- 74 Karel Capek author of "R U R" is a
 A Frenchman
 B Czechoslovakian
 C American
 D Pole
 E Italian
- 57 Lippman's *Problem of Unbelief* seeks primarily to
 A show how peace of mind can be achieved
 B denounce the creeds of the leading Protestant denominations
 C show that Christianity is a dying religion
 D determine the causes of modern man's lack of religious faith
 E improve the morals of American youth

Complete option information for high and low group responses to these 25 poorly discriminating items is provided in Table 45 where bold-faced type indicates the number of persons marking the keyed option

The 20 speech items (Numbers 26-45) make up only 20 per cent of the entire test, yet 14 of these (70 per cent) appear among the 25 least discriminating items. The first 25 items in the test cover spelling grammar and punctuation. Only 2 of these (8 per cent) appear in Table 45. The last 55 items deal with literature. 9 of these (16 per cent) were poor discriminators. It is obvious, then, that from the percentage standpoint about 5 times as many speech items as non speech ones seem unsatisfactory (70 versus 14 per cent).

There are several possible explanations for the poor showing of the speech items. First of all, they make up only a fifth of the test and therefore do not have much weight in determining each testee's total score. If speech ability is little related to the other components of the test then no matter how carefully prepared the speech items are, most of them will seem to discriminate poorly.

A second, rather likely explanation is that good speech items are quite hard to construct, while the more standard phases of English are easier to test. Some of them (Numbers 35, 38, 42, 32, 29, 45 and 44) are too easy and cannot discriminate well. Only 2 (Numbers 34 and 33) are near the optimal difficulty level of 50 per cent.

A third possibility, related to the second, is that the speech specialist was a less competent item writer than the other three staff members, or that he exercised less care than they. All four persons were experienced teachers but novices in constructing objective questions.

E. A Discrimination Table

The process of item analysis can be simplified by reference to Table 46. Knowing the number of persons tested, one immediately reads n the proper number for the high or low group. Then for the number of options each item in the test or subtest has, find the minimum value of $W_L - W_H$ needed in order to conclude that the item has significant discriminating power. In the above example, where n was 243 n is seen to be 66, and the minimum $W_L - W_H$ for acceptable discrimination with

TABLE 4C

TABLE FOR DETERMINING WHETHER OR NOT A GIVEN TEST ITEM DISCRIMINATES SIGNIFICANTLY BETWEEN A "HIGH" AND A "LOW" GROUP*

(W_L = number of persons in the low group who answered the item incorrectly or omitted it W_H = number in the high group who answered the item incorrectly or omitted it)

Total Number of Persons Tested (N)	Number in Low or High Group (0-27) ($N_L = N_H = n$)	$(W_L - W_H)$ at or above Which an Item Can Be Considered Sufficiently Discriminating			
		Number of Options			
		2 (True-False or Two-Option Multiple Choice)	3	4	5
28- 31	8	4	5	5	5
32- 35	9	5	5	5	5
36- 38	10	5	5	5	5
39- 42	11	5	5	5	5
43- 46	12	5	5	6	5
47- 49	13	5	6	6	6
50- 53	14	5	6	6	6
54- 57	15	6	6	6	6
58- 61	16	6	6	6	6
62- 64	17	6	6	6	7
65- 68	18	6	6	7	7
69- 72	19	6	7	7	7
73- 75	20	6	7	7	7
76- 79	21	6	7	7	7
80- 83	22	7	7	7	7
84- 86	23	7	7	7	7
87- 90	24	7	7	8	8
91- 94	25	7	7	8	8
95- 98	26	7	8	8	8
99- 101	27	7	8	8	8
102- 105	28	7	8	8	8
106- 109	29	7	8	8	8
110- 112	30	7	8	8	8
113- 116	31	8	8	8	8
117- 120	32	8	8	9	9
121- 124	33	8	8	9	9
125- 127	34	8	9	9	9
128- 131	35	8	9	9	9
132- 135	36	8	9	9	9
136- 138	37	8	9	9	9
139- 142	38	8	9	9	9
143- 146	39	8	9	9	9
147- 149	40	9	9	9	10
150- 153	41	9	9	10	10
154- 157	42	9	9	10	10
158- 161	43	9	10	10	10
162- 164	44	9	10	10	10
165- 168	45	9	10	10	10

TABLE 46 (Continued)

Total Number of Persons Tested (N)	Number in Low or High Group (0-27N) ($N_L = N_H = n$)	$(W_L - W_H)$ at or above Which an Item Can Be Considered Sufficiently Discriminating			
		Number of Options			
		2 (True-False or Two-Option Multiple Choice)	3	4	5
169- 172	46	9	10	10	10
173- 175	47	9	10	10	10
176- 179	48	9	10	10	10
180- 183	49	9	10	10	11
184- 187	50	9	10	11	11
188- 190	51	10	10	11	11
191- 194	52	10	10	11	11
195- 198	53	10	11	11	11
199- 201	54	10	11	11	11
202- 205	55	10	11	11	11
206- 209	56	10	11	11	11
210- 212	57	10	11	11	11
213- 216	58	10	11	11	11
217- 220	59	10	11	11	11
221- 224	60	10	11	11	12
225- 227	61	10	11	12	12
228- 231	62	10	11	12	12
232- 235	63	11	11	12	12
236- 238	64	11	11	12	12
239- 242	65	11	12	12	12
243- 246	66	11	12	12	12
247- 249	67	11	12	12	12
250- 253	68	11	12	12	12
254- 257	69	11	12	12	12
258- 261	70	11	12	12	12
262- 264	71	11	12	12	13
265- 268	72	11	12	13	13
269- 272	73	11	12	13	13
273- 275	74	11	12	13	13
276- 279	75	11	12	13	13
280- 283	76	11	12	13	13
284- 287	77	12	13	13	13
288- 290	78	12	13	13	13
291- 294	79	12	13	13	13
295- 298	80	12	13	13	13
299- 301	81	12	13	13	13
302- 305	82	12	13	13	13
306- 309	83	12	13	13	13
310- 312	84	12	13	13	14
313- 316	85	12	13	13	14
317- 320	86	12	13	14	14
321- 324	87	12	13	14	14
325- 327	88	12	13	14	14
328- 331	89	12	13	14	14
332- 335	90	12			

TABLE 46 (Continued)

Total Number of Persons Tested (N)	Number in Low or High Group ($0.27N$) ($N_L = N_H = n$)	$(W_L - W_H)$ at or above Which an Item Can Be Considered Sufficiently Discriminating			
		Number of Options			
		2 (True-False or Two-Option Multiple Choice)	3	4	5
336- 338	91	12	13	14	14
339- 342	92	12	13	14	14
343- 346	93	13	13	14	14
347- 349	94	13	14	14	14
350- 353	95	13	14	14	14
354- 357	96	13	14	14	14
358- 361	97	13	14	14	14
362- 364	98	13	14	14	14
365- 368	99	13	14	14	15
369- 372	100	13	14	14	15
406- 409	110	14	15	15	15
413- 446	120	14	15	16	16
480-483	130	15	16	16	16
517- 520	140	15	16	17	17
554- 557	150	16	17	17	18
591- 594	160	16	18	18	18
628- 631	170	17	18	19	19
665- 668	180	17	19	19	19
702- 705	190	18	19	20	20
739- 742	200	18	19	20	20
832- 835	225	19	21	21	21
925- 927	250	20	22	22	23
1017-1020	275	21	23	23	24
1110-1112	300	22	24	24	25
1480-1483	400	25	27	28	28
1850-1853	500	28	30	31	31
3702-3705	1000	39	43	44	44

* Values for this $2\frac{1}{2}$ per cent level-of-significance table, which is based upon Stanley's $\frac{1}{2}$ per cent level table (see *American Psychologist*, 6: 369, July, 1951), were computed by Miss Lilen V. Piers.

five-option items is 12⁴. Purely by accident, this makes the same number of items,

⁴ This is $\frac{1}{2} = 18$ per cent of the maximum possible $W_L - W_H$ for an n of 66. If n were 200, a difference of only 20—just 10 per cent of the maximum—would be needed for a five-option item to be significantly discriminating. With an n of 1,000, only 4.4 per cent of the maximum possible difference is required for significance. Therefore, when n is rather small, as it usually is for teacher-made tests, a considerable number of items will be branded improperly as nondiscriminating. As an extreme example, take

20, eligible for editing as were obtained by the 25 per cent rule of thumb

If a calculating machine is available $\frac{100 \times O}{2n(O-1)}$ may be locked in its keyboard and the percentage difficulty value for each item computed very rapidly by using $W_L + W_H$ as the multiplier. By hand the multiplication is likely to be tedious and inaccurate therefore in the absence of a machine it is recommended that just three points on the difficulty scale be computed in terms of $W_L + W_H$ 16 per cent for the boundary line of a very easy item 50 per cent for the middle-difficulty item and 84 per cent for the boundary line of a very hard item. The formulas for obtaining these three critical $W_L + W_H$ points are shown in Table 47.

TABLE 47

FORMULAS FOR FINDING ($W_L + W_H$) VALUES AT THREE DIFFICULTY LEVELS

Percentage of Testees Who Do Not Know the Correct Answer to the Item	Number of Options Each Item Has			
	2	3	4	5
16	$0.160n^*$	$0.213n$	$0.240n$	$0.256n$
50	$0.500n$	$0.667n$	$0.750n$	$0.800n$
84	$0.810n$	$1.120n$	$1.260n$	$1.344n$

* n = number of examinees in the low or the high group = 27 per cent of the total number tested rounded off to the nearest whole number

The three $W_L + W_H$ figures may be used to determine roughly whether the item is quite easy, of moderate difficulty or quite hard. For purposes of arranging the revised items in order of difficulty the various $W_L + W_H$ values adjusted subsequently for changes in difficulty brought about by editing will suffice.

As shown in the first footnote to Table 42 on page 438 $\frac{100 \times O}{2n(O-1)} = 0.947$ when $O = 5$ and $n = 66$. The last column of that table was obtained by multiplying each $W_L + W_H$ by 0.947.

For five options and 66 testees in each group ($n = 66$) the 16 per cent difficulty point of Table 42 occurs when $W_L + W_H = 0.256n = (0.256)(66) = 17$. Similarly the 50 per cent point is $W_L + W_H = (0.800)(66) = 53$. The 84 per cent point occurs when $W_L + W_H = (1.344)(66) = 89$. Therefore in Table 42 items will be considered easy if $W_L + W_H$ is 17 or less, of moderate difficulty if $W_L + W_H$ is from 18 to 88 and hard if $W_L + W_H$ equals 89 or more. According to this method 17 of the 100 items in Table 42 are easy while only 5 are hard. Thus the test as a whole is relatively easy.

31 examinees ($N = 31$) for whom the number in the high or the low group (n) is 8. By sheer chance marking 4 out of the 8 persons in the low group would be expected to give the keyed answer to a true-false item if none of them omitted it. Thus the item could be deemed discriminating according to Table 46 where a difference of at least 4 is required only if every person in the high group marked it correctly. Table 46 has some value as a means of determining for how many items in a test complete option information should be tabulated and a thorough scrutiny made but the basic rule illustrated in Table 42 on pages 438-440 is to edit as many items as possible beginning with the least discriminating ones.

F. Obtaining the Mean and the Standard Deviation

By using the sum of the $W_L + W_H$ column of the item-analysis table, it is rather easy to estimate the average "wrongs" score of the N testees. To secure this mean wrongs score not corrected for chance, simply add the $W_L + W_H$ column and divide this sum by $2n$

$$M_w = \frac{\Sigma(W_L + W_H)}{2n}$$

To correct the mean wrongs score for chance, use the following formula

$$M_{w_c} = \frac{O\Sigma(W_L + W_H)}{2n(O - 1)}$$

where O is the number of options each item has

The standard deviation of the wrongs scores is the same as the standard deviation of the rights scores when omits are counted as being wrong, and this statistic, not corrected for chance, is easily estimated by means of the formula

$$\sigma = \frac{\Sigma(W_L - W_H)}{2.45n}$$

Note the minus sign in this formula. To secure a standard deviation corrected for chance, multiply the above formula by $\frac{O}{(O - 1)}$

$$\sigma_c = \frac{O\Sigma(W_L - W_H)}{2.45n(O - 1)}$$

For illustrations, turn back to Table 42, page 438. There the $W_L + W_H$ column sums to 4088, and the $W_L - W_H$ column total is 1762. 4,088 divided by $2n$ equals $\frac{4088}{200}$, or 31.0, the mean number of incorrect responses, uncorrected for chance, for the 100 items.

$(5 \times 4088) - (2 \times 66 \times 4) = 38.7$ the mean wrongs corrected for chance. This figure agrees rather well with the mean of the right-hand (difficulty) column of Table 42 which is 39.0. Thus on the average the correct answer to the 100 items was "known" by about 61 per cent of the examinees and not "known" by about 39 per cent.

The standard deviation not corrected for chance is $1762 - (2.45 \times 66) = 10.9$, exactly the same as when computed from all 243 total scores. Corrected for chance it becomes $(5)(1762) - (2.45)(66)(4) = 13.6$.

These approximation formulas based upon only low and high groups are accurate enough for use in most school situations, particularly when the number of testees is fairly large, say 100 or more.

G. A Simplified Procedure for Obtaining a Reliability Coefficient

Unfortunately there is no fairly precise method for securing a single-form reliability coefficient ("coefficient of equivalence") without considerable computation. Stanley has devised two shorter procedures yielding results closely approximating those of the conventional methods. His simplified split-half technique has been reported elsewhere⁵ in considerable detail and will not be repeated here. Instead,

⁵ Julian C. Stanley, A Simplified Method for Estimating the Split-Half Reliability Coefficient of a Test, *Harvard Educational Review*, 21: 221-224, Fall 1951.

a Kuder-Richardson Formula 20 (KR_{20}) r will be obtained from just the low and high group figures in Table 42, page 438 k being the number of items

$$\begin{aligned} KR_{20} &= \frac{k}{k-1} \left\{ 1 - \frac{2n \Sigma(W_L + W_H) - \Sigma(W_L + W_H)^2}{0.667[\Sigma(W_L - W_H)]^2} \right\} \\ &= \frac{100}{99} \left[1 - \frac{2(66)(4088) - 227,630}{0.667(1762)^2} \right] \\ &= \frac{100}{99} \left(1 - \frac{311,986}{2,070,798} \right) = \frac{100}{99} (0.849) \\ &= .86, \end{aligned}$$

which is the same as the value secured by using the regular KR_{20} formula with all 243 cases. In general the abbreviated procedure yields slightly lower r 's but for most practical purposes this negative bias will be negligible.

The only part of the above formula not used in computing either the mean or the standard deviation is $\Sigma(W_L + W_H)^2$. To get it square each of the 100 ($W_L + W_H$) values in Table 42 and then sum them: $(71)^2 + (3)^2 + (77)^2 + \dots + (57)^2 = 227,630$. By hand this is a laborious process indeed but on an electric calculating machine (Friden, Marchant, Monroe) it is simple to secure both $\Sigma(W_L + W_H)$ and $\Sigma(W_L + W_H)^2$ in a single set of operations. In most medium sized and large school systems there is a machine of this sort and someone who knows how to use it. Getting all three needed values for the formula and checking them should take a skilled operator not more than half an hour, if an item analysis table similar to Table 45 has already been prepared.

This method does not involve splitting the test into halves a tedious undertaking at best when a test-scoring machine is not available. The split-half coefficient of equivalence based upon all 243 testees is .87, it is also .87 when determined from only the high and low groups. KR_{20} coefficients tend to be a little smaller than split-half ones,⁶ but again the discrepancy is usually of no practical consequence to the teacher.

⁶ Lee J. Cronbach, Coefficient Alpha and the Internal Structure of Tests, *Psychometrika* 16: 297-334, September 1951.

APPENDIX

C

Scoring Rearrangement (Ranking) Test Items

In many subjects, especially history, where one of the objectives is to acquire a sense of sequence or chronology, rearrangement questions may be a better testing device than other item forms. Their construction calls for considerable skill in putting together material so that the ranking task will not be too demanding at some points and too easy at others. Comparatively little has been written about the preparation of this type of item, though much has been said about scoring it, but in all likelihood there are potential uses for rearrangement items in many academic fields. Steps in solving a mathematical problem, sequences of equations in chemistry, and relative quality of various literary selections are possible illustrations. An essential condition is that "experts" can agree reasonably well as to how the things should be ranked, since otherwise it will not be possible to devise a satisfactory key. For this reason, independent keying by at least two teachers and reconciliation of differences *before* the items are used is desirable, but such precautions should by no means be confined to rearrangement items.

A chronology test item was presented in Table 19 on page 95 and scored for two examinees by means of the Spearman rank-difference coefficient of correlation, rho. A little reflection will make it obvious that the *magnitude* of each discrepancy between the student's response and the keyed rank, not just the fact that they differ, is taken into account. From the historian's point of view, it is far worse to think that the French Revolution occurred before the Roman Empire fell than to place it just prior to the destruction of the Spanish Armada. Similarly, it is 'wronger' to rate cork heavier than white oak than to confuse the densities of cork and balsa.

The apparent difficulty of scoring rearrangement items has discouraged most testers from using them. Actually, when a suitable table is available, the task is not formidable. One way to score each item is to compute the rho (ρ) between the student's responses and the teacher's key, as shown on page 95, and to multiply this rho by N , the number of things ranked. Thus the testee whose rankings on a certain six-option item agree completely with the key secures a rho of +1 and a score of $1 \times 6 = 6$ for the item. If his rho on another such item is 0, he gets $0 \times 6 = 0$ points credit for that item. By being completely misinformed and

securing a rho of -1 , he could earn $(-1) \times 6 = -6$ points. The rearrangement item is just about the only type that takes misinformation explicitly into account.

But using the formula, $\text{score} = (\text{number of things ranked}) \times \text{rho}$, the various values of ρ are treated as if they constituted an equal unit scale. As shown in Figure 4 on page 89, differences at the high end of the r scale represent greater changes in relationship than do differences of like magnitude between low r 's. In an attempt to compensate for these inequalities, one of the writers (JCS) prepared Table 48 by means of the formula $S = (\text{number of things ranked}) \times \text{rho squared} = N\rho^2$.

TABLE 48

ΣD^2 TABLE FOR SCORING REARRANGEMENT ITEMS
 $\text{SCORE} = (\text{NUMBER OF THINGS RANKED}) \times (\text{RHO})^2$
 (Look up ΣD^2 in Appropriate Column)

Score	Number of Things Ranked (N)						Score
	2	3	4	5	6	7	
7						0-2	7
6					0	4-6	6
5				0	2-4	8-10	5
4			0	2	6-8	12-16	4
3		0	2	4	10-12	18-22	3
2	0			6-8	14-16	24-30	2
1		2	4-6	10-12	18-24	32-40	1
0		4	8-12	14-26	26-44	42-70	0
-1		6	14-16	28-30	46-52	72-80	-1
-2	2			32-34	54-56	82-88	-2
-3		8	18	36	58-60	90-94	-3
-4			20	38	62-64	96-100	-4
-5				40	66-68	102-104	-5
-6					70	106-108	-6
-7						110-112	-7

To score a rearrangement item simply obtain the ΣD^2 "in your head" by comparing the student's responses with a conveniently arranged key. This ΣD^2 is used to enter the column of the table for the proper number of things ranked from which the score is read directly at either the left or the right. All ΣD^2 values for $N = 2$ through $N = 7$ are contained in Table 48. Tables for N 's higher than 7 can be prepared rather easily, but it does not seem wise to construct these and thereby encourage teachers to devise more complex and less homogeneous rearrangement items than are usually desirable.

Turn back to Table 19 on page 95 for two illustrations of how Table 48 is used. There Richard Roe's ΣD^2 is 40 for 6 events ranked. Looking in the "6" column of Table 48 we find that 40 lies within the 26-44 ΣD^2 range for which the score is 0. Richard gets no points at all for his inaccurate responses to this item. Had we employed the formula $S = N\rho$, he would have received $6(-14) = -84 = -1$ point.

John Doe, the other testee for whom ranks are listed in Table 19, had a ΣD^2 of 6 which lies within the 6-8 interval of Table 48 and merits a score of 4. Had he been scored by means of the $N\rho$ formula he would have obtained $83 \times 6 = 498 = 5$ points. The $N\rho^2$ scoring method of Table 48 usually yields scores closer to 0 than the $N\rho$ procedure. It seems somewhat more defensible on statistical grounds.

APPENDIX

D

The Computation of Square Roots

When determining a standard deviation or an r , one must find the square root of a number. This can be done easily from a table of square roots, such as Barlow's *Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals* (London: E. & F. N. Spon, Ltd.). Quite a few statistics texts and other books have shorter tables. If only an occasional square root is needed, it can probably be secured more easily 'by hand' than by searching for a table.

In Chapter 15 of *Mathematics Essential for Elementary Statistics* (New York: Henry Holt and Company, 1951), Helen M. Walker devotes 16 pages to explaining what square roots mean and how they are secured. The reader who is thoroughly in the dark concerning this topic will want to consult that reference. For others who need merely a little reviewing of material previously learned, the following brief explanation is offered.

1. First take a small three-digit whole number that is a perfect square, 144 is a good illustration. The square root of 144 is a number which, when multiplied by itself, equals 144. To extract the square root of 144 follow these seven steps:

$$(a) \quad \sqrt{144}$$

$$(b) \quad \sqrt{1 \wedge 44}$$

$$(c) \quad \begin{array}{r} \sqrt{1 \wedge 44} \\ -1 \\ \hline 0 \end{array}$$

$$(d) \quad \begin{array}{r} \sqrt{1 \wedge 44} \\ -1 \\ \hline 0 \quad 44 \end{array}$$

$$\begin{array}{r}
 (e) \quad \sqrt{1 \text{ } 44} \\
 \underline{-1} \\
 2 \overline{) 0 \ 44}
 \end{array}$$

$$\begin{array}{r}
 (f) \quad \sqrt{1 \text{ } 2 \text{ } 44} \\
 \underline{-1} \\
 22 \overline{) 0 \ 44}
 \end{array}$$

$$\begin{array}{r}
 (g) \quad \sqrt{1 \text{ } 2 \text{ } 44} \\
 \underline{-1} \\
 22 \overline{) 0 \ 44} \\
 \underline{-44} \\
 00
 \end{array}$$

- (a) Write the 144 with a square root sign and two decimal points one above the other
- (b) Begin with the decimal point following 144 and move to the left two digits at a time, putting a caret at each stopping place. With 144 only one move and therefore one caret is needed
- (c) Look at the number to the left of the caret. What number multiplied by itself is as nearly equal to 1 as possible but not greater than 1? 1, of course so write this 1 above the 1 and below it. Subtract
- (d) Draw down the next two numbers
- (e) Double the top 1 and write it to the left of 0 44
- (f) Now, how many times does 2 go into 0 44? 2 so write 2 in the answer space above the right-hand 4 and also to the right of the 2
- (g) Multiply the 22 by 2 write this product (44) below the other 44 and subtract. Therefore, the square root of 144 is exactly 12 since $12 \times 12 = 144$
2. Take a large decimal fraction 9342 156, and find its square root to the nearest two decimal places

$$(a) \quad \sqrt{93 \text{ } 42 \text{ } 15 \text{ } 60 \text{ } 00}$$

$$\begin{array}{r}
 (b) \quad \sqrt{9 \text{ } 34 \text{ } 21 \text{ } 56 \text{ } 00} \\
 \underline{-81} \\
 12
 \end{array}$$

$$\begin{array}{r}
 (c) \quad \sqrt{9 \text{ } 34 \text{ } 21 \text{ } 56 \text{ } 00} \\
 \underline{-81} \\
 18 \overline{) 12 \ 42}
 \end{array}$$

$$\begin{array}{r}
 (d) \quad \sqrt{9 \text{ } 6 \text{ } 34 \text{ } 21 \text{ } 56 \text{ } 00} \\
 \underline{-81} \\
 186 \overline{) 12 \ 42} \\
 \underline{-11 \ 16} \\
 1 \ 26
 \end{array}$$

$$\begin{array}{r}
 966 \\
 \sqrt{93\wedge 4215\wedge 60\wedge 00\wedge} \\
 \underline{-81} \\
 1861242 \\
 \underline{-1116} \\
 192612615 \\
 \underline{-11556} \\
 1059
 \end{array}$$

$$\begin{array}{r}
 96654 \\
 \sqrt{93\wedge 4215\wedge 60\wedge 00\wedge} = 9665 \\
 \underline{-81} \\
 1861242 \\
 \underline{-1116} \\
 192612615 \\
 \underline{-11556} \\
 19325105960 \\
 \underline{-96625} \\
 193304933500 \\
 \underline{-773216} \\
 160284
 \end{array}$$

- (a) First write down the number with the square root sign, carets, and a decimal point in the answer place (Notice that this decimal point is always exactly above the decimal point in the number) Begin at the decimal point in the number and count in both directions by two's putting a caret between each pair. Zeros are added to the right of the decimal point beyond the last figure in order to have the two numbers to draw down each time. In order to carry out the square root to the nearest two decimal places (rounded off from three places) it is necessary to have six figures to the right of the decimal point.
- (b) What number multiplied by itself is as nearly equal to 93 as possible, without exceeding it? $10 \times 10 = 100$ which is too much. $9 \times 9 = 81$, so use 9 as the first number in the square root. Multiply it by itself and subtract the 81 from 93.
- (c) Draw down the next two numbers (42), double 9, and write 18 to the left of 12 42.
- (d) Approximately how many times will 18 go into 124? Not quite 7, for $18 \times 7 = 126$. Try the next lower number, 6. Write it in the answer space above the 2 and also to the right of the 18. Multiply 6 by 186 and subtract this product, 11 16 from 12 42.
- (e) Draw down the next two numbers and double the 96. 19 goes into 126 about 6 times. Repeat the above process.
- (f) Double 966 write 1932 in the proper place, and complete the remaining steps. The square root of 9342 156 is 96 654 +, which when rounded off to the nearest two decimal places becomes 96 65. Where test scores are concerned, only one decimal place is usually needed for the standard deviation. Also, in the denominator of the r formula on page 92 extraction of the square roots to the nearest three figures is sufficient for most purposes.

To check the computation of a square root, multiply the value obtained by itself and add to this product the remainder. For example $(96\ 654 \times 96\ 654) + 160284 = 9342\ 156000$ which agrees exactly with the figure underneath the square root sign in Step (a) on page 457.

APPENDIX

E

Answers to Questions in Appendix A

- 1 C. "The sum of f " is the total frequency N
- 2 D. For example 310-319 means 310 311 312 313 314, 315 316 317 318 and 319—a total of 10 numbers Likewise the difference between the upper and lower "real" class limits = $(319 + 0.5) - (310 - 0.5) = 319.5 - 309.5 = 10$
- 3 A. See above
- 4 B. $(260 + 269)/2 = 529/2 = 264.5$ Likewise $259.5 + (10/2) = 269.5 - (10/2) = 264.5$
- 5 B. The assumed mean lies at the midpoint of the class, 260-269 whose d is 0
Thus $M = (260 + 269)/2 + [10 \times (-15)]/50 = 264.5 + (-150)/50 = 264.5 - 3 = 261.5$
- 6 C. $50/2 = 25$ $259.5 + \left(\frac{25 - 24}{12}\right)(10) = 259.5 + (10/12) = 260.3$ Similarly, counting from the top down as a check, $269.5 - \left(\frac{25 - 14}{12}\right)(10) = 269.5 - (110/12) = 269.5 - 9.2 = 260.3$
- 7 D. The mode is the midpoint of the class having the greatest f 12 is the largest figure in the f column and the midpoint of its class is $(260 + 269)/2 = 264.5$
- 8 D. $\frac{1}{4}$ of 50 is 12.5 Count up 12.5 frequencies $239.5 + \left(\frac{12.5 - 5}{8}\right)(10) = 239.5 + (75/8) = 239.5 + 9.4 = 248.9$ Or, $\frac{3}{4}$ of 50 = $150/4 = 37.5$ Counting down, $249.5 - \left(\frac{37.5 - 37}{8}\right)(10) = 249.5 - (.6) = 249.5 - 0.6 = 248.9$
- 9 D. The 75th percentile is $\frac{3}{4}$ of the way from the bottom of the distribution and $\frac{1}{4}$ of the way from the top $\frac{1}{4}$ of 50 is 12.5 Count down 12.5 frequencies $279.5 - \left(\frac{12.5 - 8}{6}\right)(10) = 279.5 - (45/6) = 279.5 - 7.5 = 272.0$ To check, count up 37.5 frequencies $269.5 + \left(\frac{37.5 - 36}{6}\right)(10) = 269.5 + (15/6) = 269.5 + 2.5 = 272.0$

- 10 B. Q_1 is the 75th percentile, found to be 272.0 in Question 9, above, and Q_3 is the 25th percentile, 248.9 in Question 8 $(272.0 - 248.9)/2 = (23.1)/2 = 11.55 = 12$
- 11 C. $\frac{1}{8}$ of 50 is 5 $239.5 + \left(\frac{5-5}{8}\right)(10) = 239.5$ Counting down, $\frac{3}{8}$ of 50 is 45 $239.5 - \left(\frac{45-45}{2}\right)(10) = 239.5$
- 12 C. $50 - 0.9(50) = 5$ Count down 5 or count up 45 Obviously, it is easier to count down 5 Counting up is useful as a check, though $299.5 - \left(\frac{5-3}{4}\right)(10) = 299.5 - (20/4) = 294.5$ Thus the 90th percentile is the midpoint of the 290-299 class it lies exactly halfway within that class, 5 units below the upper real limit and 5 units above the lower real limit Check by counting up $289.5 + \left(\frac{45-43}{4}\right)(10) = 289.5 + 5 = 294.5$
- 13 B Use the answers to Questions 12 and 11 in solving this $0.4(294.5 - 239.5) = 0.4(55) = 22.0$
- 14 F. $10 \times \frac{\sqrt{50(239) - (-15)^2}}{50} = \frac{\sqrt{11,950 - 225}}{5} = \frac{\sqrt{11,725}}{5}$
 $= \frac{108}{5} - 21.6 = 22$
- 15 B. $19.5 - 9.5 = 10$
- 16 C. $99.5 + \frac{1}{2}(109.5 - 99.5) = 99.5 + \frac{1}{2}(10) = 104.5$ To check $109.5 - \frac{1}{2}(109.5 - 99.5) = 109.5 - 5 = 104.5$
- 17 A. There are two essentially different reasons for grouping scores computational (to reduce labor) and graphical (to emphasize important features of the data) One of the best discussions of the latter aspect is contained in Truman L. Kelley's *Fundamentals of Statistics*, Chapter IV, "Graphic Methods" Cambridge, Massachusetts Harvard University Press, 1947
- 18 A It is necessary to know the range *before* performing the operations set forth in Options B, C, D, and E
- 19 E. For instance, into which of the two classes, 44-48 and 40-44, would you put a score of 44?
- 20 D. The 60th percentile is defined as the point in a distribution below which lie 60 per cent of the scores and above which lie 40 per cent of the scores
- 21 D. When arranged in numerical order, these scores are 4 4 5 6, 7 The middle score (mid-score) is 5
- 22 C. In numerical order these scores are 44, 46, 46, 46, 63, 68, 68, their mid-score is 46, which has three numbers on each side of it A frequency distribution of these scores with an interval of 1 is as follows:

Score	f
68	2
63	1
46	3
44	1

The median of this distribution is found by counting half the way up or down the frequency column The total frequency is 7, and half of 7 is 3.5 Counting up through the 43.5-44.5 class uses only 1 frequency, but count-

ing through the next (45.5-46.5) class involves $1 + 3 = 4$ frequencies, more than the 3.5 required to locate the median. Thus the median is $45.5 + \left(\frac{3.5 - 1}{3}\right)(1) = 45.5 + (2.5)/3 = 45.5 + 0.83 = 46.33$. To check $46.5 - \left(\frac{3.5 - 3}{3}\right)(1) = 46.5 - (0.5/3) = 46.5 - 0.17 = 46.33$. Therefore, the discrepancy between the median and the midscore in this distribution is $46.33 - 46 = 0.33$, which illustrates the fact that the midscore and the median of a distribution may have different values. Usually the difference is slight, however.

23. C. $(4 + 5 + 7 + 6 + 4)/5 = 26/5 = 5.2$

24. C. Only two of the five options (C and E) contain measures of central tendency. The standard deviation, quartile deviation and range are measures of variability. Since the arithmetic mean is a function of every score in the distribution, its value would reflect "the undue influence of a few extreme values." Whether the highest paid worker made \$5000 or \$50,000 is wholly inconsequential so far as the size of the median is concerned.

25. E.

26. D.

27. C. $M = M + \frac{1 \times \Sigma fd}{N} = (0 + 3)/2 + \frac{[3.5 - (-0.5)] \times [(7 \times 0) + (3 \times 1)]}{7 + 3}$
 $= 1.5 + \frac{4 \times 3}{10} = 1.5 + 1.2 = 2.7$

28. A. It shows the range of the middle 80 per cent of the scores in the distribution.

29. A. Q_3 is the 75th percentile, and the median is the 50th percentile. Twenty-five per cent of the scores lie between these two points, since $75 - 50 = 25$.

30. E.	Score	f	d	fd	fd'
	4	3	2	6	12
	3	0	1	0	0
	2	4	0	0	0
	1	0	-1	0	0
	0	3	-2	-6	12
		<hr/>	<hr/>	<hr/>	<hr/>
		N = 10		$\Sigma fd = 0$	$\Sigma fd' = 24$

$$SD = \frac{1 \times \sqrt{N \Sigma fd'^2 - (\Sigma fd')^2}}{N} = \frac{1 \times \sqrt{10(24) - 0^2}}{10}$$

$$= \frac{\sqrt{240}}{10} = \frac{15.5}{10} = 1.55$$

$$\begin{array}{r} 1 \ 5 \ 5 \\ \sqrt{2 \ 4 \ 0 \ 0} \\ 1 \end{array}$$

$$\begin{array}{r} 25 \ 1 \ 40 \\ 1 \ 25 \end{array}$$

$$\begin{array}{r} 305 \ 1500 \\ 1525 \\ \hline -25 \end{array}$$

- 10 B Q_2 is the 75th percentile, found to be 272.0 in Question 9, above, and Q_1 is the 25th percentile, 248.9 in Question 8 $(272.0 - 248.9)/2 = (23.1)/2 = 11.55 = 12$
- 11 C $\frac{1}{10}$ of 50 is 5 $239.5 + \left(\frac{5-5}{8}\right)(10) = 239.5$ Counting down, $\frac{1}{10}$ of 50 is 45 $239.5 - \left(\frac{45-45}{2}\right)(10) = 239.5$
- 12 C $50 - 0.9(50) = 5$ Count down 5 or count up 45 Obviously, it is easier to count down 5 Counting up is useful as a check, though $299.5 - \left(\frac{5-3}{4}\right)(10) = 299.5 - (20/4) = 294.5$ Thus the 90th percentile is the midpoint of the 290-299 class it lies exactly halfway within that class, 5 units below the upper real limit and 5 units above the lower real limit Check by counting up $289.5 + \left(\frac{45-43}{4}\right)(10) = 289.5 + 5 = 294.5$
- 13 B Use the answers to Questions 12 and 11 in solving this $0.4(294.5 - 239.5) = 0.4(55) = 22.0$
- 14 F $10 \times \frac{\sqrt{50(239) - (-15)^2}}{50} = \frac{\sqrt{11,950 - 225}}{5} = \frac{\sqrt{11,725}}{5}$
 $= \frac{108}{5} = 21.6 = 22$
- 15 B $19.5 - 9.5 = 10$
- 16 C $99.5 + \frac{1}{2}(109.5 - 99.5) = 99.5 + \frac{1}{2}(10) = 104.5$ To check $109.5 - \frac{1}{2}(109.5 - 99.5) = 109.5 - 5 = 104.5$
- 17 A There are two essentially different reasons for grouping scores: computational (to reduce labor) and graphical (to emphasize important features of the data) One of the best discussions of the latter aspect is contained in Truman L. Kelley's *Fundamentals of Statistics*, Chapter IV, "Graphic Methods" Cambridge, Massachusetts: Harvard University Press, 1947
- 18 A It is necessary to know the range *before* performing the operations set forth in Options B, C, D, and E
- 19 E For instance, into which of the two classes, 44-48 and 40-44, would you put a score of 44?
- 20 D The 60th percentile is defined as the point in a distribution below which lie 60 per cent of the scores and above which lie 40 per cent of the scores
- 21 D When arranged in numerical order, these scores are 4, 4, 5, 6, 7 The middle score (mid-score) is 5
- 22 C In numerical order these scores are 44, 46, 46, 46, 63, 68, 68, their mid-score is 46 which has three numbers on each side of it A frequency distribution of these scores with an interval of 1 is as follows:

Score	f
68	2
63	1
46	3
44	1

The median of this distribution is found by counting half the way up or down the frequency column The total frequency is 7, and half of 7 is 3.5 Counting up through the 43.5-44.5 class uses only 1 frequency but count-

ing through the next (45.5-46.5) class involves $1 + 3 = 4$ frequencies, more than the 3.5 required to locate the median. Thus the median is $45.5 + \left(\frac{3.5 - 1}{3}\right)(1) = 45.5 + (2.5)/3 = 45.5 + 0.83 = 46.33$. To check $46.5 - \left(\frac{3.5 - 3}{3}\right)(1) = 46.5 - (0.5/3) = 46.5 - 0.17 = 46.33$. Therefore, the discrepancy between the median and the midscore in this distribution is $46.33 - 46 = 0.33$, which illustrates the fact that the midscore and the median of a distribution may have different values. Usually the difference is slight, however.

23. C. $(4 + 5 + 7 + 6 + 4)/5 = 26/5 = 5.2$

24. C. Only two of the five options (C and E) contain measures of central tendency. The standard deviation, quartile deviation, and range are measures of variability. Since the arithmetic mean is a function of every score in the distribution, its value would reflect "the undue influence of a few extreme salaries." Whether the highest paid worker made \$60,000 or \$50,000 is wholly inconsequential so far as the size of the median is concerned.

25. F.

26. D.

27. C. $M = M + \frac{1 \times \Sigma fd}{N} = (0 + 3)/2 + \frac{[3.5 - (-0.5)] \times [(7 \times 0) + (3 \times 1)]}{7 + 3}$
 $= 1.5 + \frac{4 \times 3}{10} = 1.5 + 1.2 = 2.7$

28. A. It shows the range of the middle 80 per cent of the scores in the distribution.

29. A. Q_3 is the 75th percentile, and the median is the 50th percentile. Twenty-five per cent of the scores lie between these two points, since $75 - 50 = 25$.

30. E.	Score	f	d	fd	fd ²
	4	3	2	6	12
	3	0	1	0	0
	2	4	0	0	0
	1	0	-1	0	0
	0	3	-2	-6	12
		<hr/>	<hr/>	<hr/>	<hr/>
		$N = 10$		$\Sigma fd = 0$	$\Sigma fd^2 = 24$

$$SD = \frac{1 \times \sqrt{N \Sigma fd^2 - (\Sigma fd)^2}}{N} = \frac{1 \times \sqrt{10(24) - 0^2}}{10}$$

$$= \frac{\sqrt{240}}{10} = \frac{15.5}{10} = 1.55$$

$$\begin{array}{r} 1 \ 5 \ 5 \\ \sqrt{240} \ 40 \ 00 \\ 1 \ \underline{\hspace{1cm}} \\ 25 \ 1 \ 40 \\ 1 \ \underline{\hspace{1cm}} \\ 305 \ 1 \ 500 \\ 1525 \ \underline{\hspace{1cm}} \\ -25 \end{array}$$

The size of the interval of each class is 1, the classes 2.5-3.5 and 0.5-1.5 having frequencies of 0. For convenience the assumed mean M' , was at 2, the midpoint of the 1.5-2.5 class, which is the center of the distribution. It may be put anywhere else, of course, without altering the answer.

$$\begin{aligned}
 31 \text{ D. } Q &= \frac{75\text{th percentile} - 25\text{th percentile}}{2} \\
 &= \frac{\left[5.5 - \left(\frac{4-3}{4} \right) (1) \right] - \left[1.5 + \left(\frac{4-2}{3} \right) (1) \right]}{2} \\
 &= \frac{(5.5 - 0.25) - (1.5 + 0.67)}{2} = \frac{5.25 - 2.17}{2} = \frac{3.08}{2} = 1.54
 \end{aligned}$$

$$32 \text{ C } 2.5 + \left(\frac{8-5}{4} \right) (1) = 2.5 + 0.75 = 3.25$$

$$\text{Check } 3.5 - \left(\frac{8-7}{4} \right) (1) = 3.5 - 0.25 = 3.25$$

$$33 \text{ B } z = \frac{\text{Score} - \text{Mean}}{SD} = \frac{70 - 80}{20} = \frac{-10}{20} = -0.5$$

Therefore, this individual is half a standard deviation below the mean of the group with which he was tested.

- 34 E. Three tied scores of 95 occur. If there were no ties these three places would have ranks of 4, 5, and 6. Since one score of 95 is as good as another, we assign the average of 4, 5, and 6—which is 5—to each of the three scores. Note that $4 + 5 + 6 = 15$, the same as the sum of the new ranks $5 + 5 + 5 = 15$. Whether or not ties occur, the sum of a certain number (N) of consecutive ranks beginning with 1 will always be $[N(N+1)]/2$. If there are 9 ranks as in this question, their sum will be $(9 \times 10)/2 = 45$.

- 35 B. As noted above with reference to Question 34, the method of assigning to tied scores the average rank that would have occurred without ties keeps the sum of untied and tied sets of ranks of the same length identical. Since the sum is unchanged, the mean—which is the sum divided by N , the number of ranks, is also unchanged. It will always be $\frac{N+1}{2}$.

- 36 B. See Chapter 3.

- 37 A.

- 38 C. An r of 0 has the least possible predictive value. The closer to 0 r gets, regardless of sign, the poorer prediction becomes.

- 39 E. The r discussed in Chapter 3 simply cannot be greater than +1.00 or -1.00 except when computational errors are made.

- 40 A. Each tally mark represents a pair of scores. There are V pairs of scores in all. The number of cells in a 12×12 scatter diagram is 144, but some of these will probably be blank, while others will have more than one tally. See Table 18, page 92, which has $15 \times 14 = 210$ cells, 33 of which contain the $N = 43$ tallies. $210 - 33 = 177$ of the cells are empty.

- 41 B. The arithmetic mean is a measure of central tendency, the standard deviation is a measure of variability.

- 42 C. Q_3 is the 75th percentile, the median (Q) is the 50th percentile.

- 43 E. Both the mean and the standard deviation are based upon all scores in the distribution, the median and Q are both percentile measures. Also, the mean is used with the SD, while the median is used with Q . The analogy is: A certain kind of measure of central tendency is to a similar sort of measure of variability as another kind of measure of central tendency is to a similar sort of measure of variability.
- 44 E. A frequency distribution has a median but usually no midscore while ungrouped measures have a midscore (the middle score if the number of scores is odd, or the average of the two middle scores if the number is even).
- 45 D. 50 per cent of all the measures in a distribution always lie between Q_2 and Q_4 . In a normal (so-called "bell-shaped") distribution, 68 per cent of all cases lie within one standard deviation of the mean.
- 46 A. The arithmetic mean is the most reliable measure of central tendency, the mode the least reliable, the standard deviation is the most reliable measure of variability, the range the least reliable.
- 47 B. The standard deviation is a linear distance along the base line of a frequency distribution.
- 48 E. When correlation is positive, high scores on one test tend to go with high scores on the other test, while low scores tend to go with low scores. This is a direct relationship. When correlation is negative, high scores on one test go with low scores on the other, and vice versa. This is an inverse relationship.
- 49 A. Spearman derived the formula for the rank-difference coefficient of correlation, ρ , while somewhat earlier Pearson had derived the basic formula for r .
- 50 B. Ranks denote order only. We do not know how high or low a score the person who obtained a certain rank may have had. Scores tell how many points the testee earned—that is, the magnitude of his achievement.

APPENDIX

F

Publishers of Standardized Tests

The following list includes every test company for whom five or more tests are indexed on pages 1100-1106 of Oscar K. Buros (Editor), *The Fourth Mental Measurements Yearbook*, Highland Park, New Jersey: Gryphon Press, 1953.

The number of tests covered in the 1953 yearbook is shown in bold-face type following each address. An asterisk (*) preceding the name indicates that the company issues catalogues devoted entirely or in large part to tests" (Buros, page 1100).

- *Acorn Publishing Co., Inc., Rockville Centre, New York (21)
- *Australian Council for Educational Research, 147 Collins St., Melbourne, C 1, Australia (15)
- Benton Review Publishing Co., Inc., Fowler, Indiana (11)
- *Bureau of Educational Measurements, Kansas State Teachers College of Emporia, Emporia, Kansas (19)
- *Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa (14)
- *Bureau of Publications, Teachers College, Columbia University, New York 27, New York (11)
- *California Test Bureau, 5916 Hollywood Blvd., Los Angeles 28, California (29)
- *Center for Psychological Service, George Washington University, Washington, D. C. (5)
- College Entrance Examination Board, 425 W. 117th Street, New York 27, New York (18)
- *Cooperative Test Division, Educational Testing Service, Princeton, New Jersey (59)
- Division of Educational Reference, Purdue University, Lafayette, Indiana (6)
- Educational Records Bureau, 21 Audubon Ave., New York 32, New York (10)
- *Educational Test Bureau, Educational Publishers, Inc., 720 Washington Ave., S. E., Minneapolis, Minnesota (30)
- Educational Testing Service, Princeton, New Jersey, (19)

- **C A Gregory Co*, 345 Calhoun St, Cincinnati 19, Ohio (5)
- **George G Harrap & Co, Ltd*, 182 High Holborn, London W C 1, England (11)
- **Houghton Mifflin Co*, 2 Park St, Boston 7, Massachusetts (14)
- Joint Committee on Tests*, 132 W Chelton Ave, Philadelphia 44, Pennsylvania (7)
- National League of Nursing Education, Inc*, 2 Park Ave, New York 16, New York (5)
- **Ohio Scholarship Tests*, Ohio State Department of Education, Columbus, Ohio (29)
- Personnel Research Institute, Western Reserve University* Cleveland Ohio (9)
- **Psychological Corporation*, 522 Fifth Ave New York 18, New York (54)
- Psychological Service Center Press*, 1275 New Hampshire Ave, N W, Washington 6, D C (6)
- Psychometric Affiliates* Box 1625, Chicago 90 Illinois (5)
- **Public School Publishing Company* 509-513 North East St, Bloomington, Illinois (12)
- **Science Research Associates, Inc*, 57 West Grand Ave, Chicago 10, Illinois (37)
- **Sheridan Supply Co*, P O Box 837, Beverly Hills, California (9)
- Turner E Smith & Co*, 441 West Peachtree St, N E, Atlanta 3, Georgia (8)
- **Stanford University Press*, Stanford, California (13)
- State High School Testing Service for Indiana*, Purdue University, Lafayette, Indiana (33)
- Steck Co*, Austin 1, Texas (6)
- **C H Stoelting Co*, 424 North Homan Ave, Chicago 24, Illinois (5)
- **University of London Press, Ltd*, Little Paul's House Warwick Square, London E C 4, England (12)
- **Vocational Guidance Centre*, 371 Bloor St, W, Toronto 5 Canada (12)
- **World Book Company*, 313 Park Hill Ave, Yonkers-on Hudson 5, New York (47)

Author Index

- Adams, Eunice, 294
 Adams, Jessie E., 363
 Adkins, Dorothy C., 55, 162, 191, 245
 Allen, Mildred M., 225
 Allport, Gordon W., 49, 176, 190, 214, 421
 Anastasi, Anne, 219, 318
 Anderson, C. J., 305
 Anderson, Gordon V., 371
 Anderson, Harold A., 195, 288
 Anderson, Scarvia B., 152, 326
 Arbuckle, Dagald S., 371
 Aristotle, 8, 9, 14, 20
 Arkin, Hubert, 273
 Arnold, Dwight L., 143
 Arthur, Grace, 422
 Ashbaugh, E. J., 41, 123
 Ashburn, Robert R., 193
 Asher, E. J., 277
 Avent, Joseph E., 139
 Ayres, Leonard P., 38, 39, 45, 115
- Bain, Alexander, 52
 Baker, Arthur O., 168
 Baker, Harry J., 330, 333, 349
 Bamberger, Sister Clara Francis, 254
 Barnes, Harry Elmer, 10
 Barr, Arvil S., 15, 21, 25, 127, 378, 421
 Barrett, E. R., 182
 Barton, W. A., Jr., 174
 Bass, Bernard M., 421
 Bayless, Ernest C., 174
 Beard, Charles A., 4
 Beauchamp, W. L., 168, 171, 177
 Bebell, Clifford, 366
 Beck, Roland L., 171
 Bodell, Ralph C., 174
 Beggs, V. L., 407
 Benjamin, A. Cornelius, 14
 Bennett, George K., 134, 245, 371, 419
 Benson, Arthur L., 398
 Berdie, Ralph F., 219
 Berkshire, James H., 365
 Betts, Emmett Albert, 345
 Betts, Gilbert L., 182
 Billett, Roy O., 348, 353, 354, 358, 362
 Billings, Josh, 370
 Binet, Alfred, 30, 31, 32, 33, 34, 35, 48, 51,
 52, 53, 55, 57, 108, 109, 115, 127, 215,
 227, 229, 280, 282, 284, 290, 351, 418
- Bingham, Walter V., 350
 Bixler, Harold H., 130
 Black, Max, 14
 Blair, Glenn Myers, 331, 334, 345
 Bledsoe, Joseph C., 371
 Blin, 32
 Blommers, Paul, 22
 Bobbitt, Joseph M., 341
 Book, William F., 321
 Bordin, Edward, 144
 Borgersrode, Fred von, 221-224
 Boring, Edwin G., 6, 15, 23, 32
 Boss, Mabel E., 406
 Bowles, Frank H., 194
 Boyd, Gertrude, 345
 Boyer, Phillip A., 213, 361, 406
 Boynton, Paul L., 59, 229, 281
 Brandenburg, G. C., 49
 Brenner, Benjamin, 325, 326
 Bridges, Claude F., 246
 Brinton, W. C., 247, 271, 273
 Brown, Clara M., 183
 Brown, Edwin J., 402
 Brown, F. W., 187
 Brown, Francis J., 316, 317
 Brown, G. L., 284
 Brown, Judson S., 327
 Brown, William, 123, 124
 Brown, Woodrow A., 246
 Browne, Arthur D., 398
 Brownell, William A., 14, 126, 134, 142
 167, 224, 329, 341
 Brubacher, John S., 15, 25
 Brueckner, Leo J., 126, 336
 Bruner, Herbert B., 388
 Buckingham, B. R., 15, 43, 44
 Burbank, Luther, 299
 Burnham, Paul Sylvester, 371
 Burns, Bob, 150
 Burnside, Carolyn J., 246
 Buros, Oscar, 54, 55, 120, 121, 134, 197
 219, 245, 464
 Burt, Cyril, 31
 Buswell, Guy T., 340
- Cadwell, Dorothy H. B., 245
 Cady, 49
 Caldwell, Otis W., 29, 38
 Caldwell V. V., 354, 355

- Calvin, Allen, 51
 Cameron Dale C., 341
 Camp, 309
 Campbell Pera, 340
 Cane, V R, 327
 Cantril Hadley, 52
 Carmichael, Leonard, 47
 Carroll, John B., 135, 219
 Carter, Ralph E., 198, 200
 Cattell, J McKeen, 30, 33 52, 115
 Cattell, Psyche, 284, 285, 288
 Cattell, R B, 52
 Chadwick, E., 38
 Chalmers, T M, 314
 Chambers, E G, 104
 Charters W W, 337
 Chase, W P, 326
 Chauncey, Henry, 300, 371
 Chave, Ernest J, 59
 Childers Leon M, 41
 Clark, Champ, 19
 Clark, Edward L., 124, 348
 Clark, Willis W, 175
 Cobb, E B, 245, 371
 Cobb, Irvin S., 19
 Cochran Roy E., 194, 196, 203, 204
 Cohen, I Bernard, 25
 Cole, Robert D., 221-224
 Coleman Wilham, 245, 371
 Colton, Raymond R., 273
 Compton, R K, 321
 Conant, James Bryant, 5
 Conklin, Edmund S., 352
 Conneau, A., 163
 Conner, 310
 Conrad Herbert S., 51, 52, 120, 297
 Cook, Stuart W., 25, 51, 398, 425
 Cook, Walter W., 59, 166, 206, 300, 327, 345, 365
 Coombs Clyde H., 424
 Cooper, D H, 398
 Cornell, Ethel L., 359
 Cornell, Francis G., 394, 395, 396, 398
 Cornog, J., 168
 Courtis, Stuart A., 29, 38, 134, 219
 Cowdery, Karl M., 51
 Crawford, Albert Beecher, 371
 Crawford, John R., 291
 Crespi, Leo P., 52
 Cressey, Paul F., 27
 Cronbach, Lee J., 23, 55, 124, 154, 162, 180, 245, 300, 345, 416, 423, 453
 Crow, Alice, 365
 Crow, Lester D., 365
 Crowder, Norman A., 418
 Cureton, Edward E., 135, 298, 417, 420, 425
 Curtis, F D., 174
 Dailey, John T., 398
 Daly, Joseph F., 102
 Danielson, Paul J., 372
 Darley, John G., 278, 371, 415
 Darling, W C, 174
 Darwin, Charles, 8
 Davenport, K S., 116
 Davis, Allison, 278, 422
 Davis, Frederick B., 22, 25, 55, 120, 134, 135, 154, 157, 160, 162, 371, 422, 440
 Davis, Georgia, 341, 342
 Davis, R., 178
 Davis, Robert A., 25, 139, 164, 197, 348, 421
 Deming, W Edwards, 60
 Deputy, E C, 317, 319
 Deutsch, Morton, 25, 51, 398, 425
 Dewey, John, 4, 11, 25, 357
 Diamond, Leon N., 119
 Diederich, Paul B., 238
 Dixon, Wilfred J., 105
 Dolbear, Amos E., 352
 Dumas, Simeon J., 398, 418
 Doppelt, Jerome E., 418, 425
 Doughton, Isaac, 16
 Douglass, Earl R., 321
 Dressel, Paul L., 120, 134, 140, 345
 DuBos, Philip H., 134
 Dunlap, Jack W., 234
 Dunlap, Knight, 37
 Dunn, Leslie C., 8
 Durant, Will, 16
 Durost, Walter N., 246, 300
 Durrell, Donald D., 334
 Dyer, Henry S., 326
 Dykema, Karl W., 119
 Eaton, Merrill T., 21
 Ebbinghaus, Hermann, 30, 31, 34
 Ebel, Robert L., 191
 Edmiston, R W., 201
 Edwards, Allen L., 105
 Edwards, I Newton, 347
 Eells, Kenneth W., 278, 422
 Eells, Walter Crosby, 414
 Einstein, Albert, 7, 25
 Elliott, Edward C., 40
 Ellis, Albert, 51, 206
 Ellis, Robert S., 365
 Ellis, Robert W., 398
 Ellwood, Charles A., 10
 Flsbree, Willard S., 408, 415
 Elwell, Mary, 126, 336
 Engel Thelburn L., 363
 Engelhardt, N L Jr., 397
 Engelhardt, N L, Sr., 395
 Engelhart, Max D., 49, 134, 165
 Erickson, Clifford E., 371
 Eulich, Alvin C., 165, 311, 377, 398
 Evans, Robert O., 407, 415
 Eysenck, Hans J., 52
 Ezell, L B, 191
 Fabre, Jean Henri, 9
 Falls J D., 40
 Fan Chung-Teh, 134
 Farley, Belmont Mercer, 402, 403
 Fechner, Gustav T., 31
 Ferguson, George A., 125
 Fernald, G G., 47
 Fernald, Grace M., 345
 Ferrell, Guy V., 88

- Fida, Silvio, 7
 Finch, Frank H., 182, 238
 Findley, Warren G., 134
 Fish, Louis J., 405
 Fisher, George, 38
 Fisher, Ronald A., 8, 88
 Flanagan, John C., 52, 134, 146, 204, 206, 245, 274, 288, 300
 Fletcher, Marie A., 364
 Foley, J. D., 372
 Foley, John P., Jr., 348
 Folk, S. B., 361
 Forlino, George, 316
 Fowler, Fred M., 369
 Franzen, Raymond, 298
 Frederiksen, Norman, 246, 300, 371
 Freeman, Frank N., 59
 Freeman, Frank S., 55, 135, 352
 Froelich, Clifford P., 371, 415
 Furfey, Paul H., 102

 Gable, Sister Felicitia, 310
 Gage, N. L., 206
 Galen, 17
 Galileo, 4, 6, 7, 11
 Gallup, George H., 52
 Galton, Sir Francis, 8, 30, 31, 33, 38, 46, 48, 49, 52, 85, 86, 109, 115, 308, 348
 Gardner, Eric F., 59, 162, 300
 Garrett, Harley I., 86
 Garrett, Henry E., 105, 290
 Gates, Arthur I., 182, 357
 Gerberich, Raymond J., 22
 Gilbert, Arthur W., 365
 Gilchrist, Robert S., 415
 Gillespie, F. H., 49, 50
 Gilliland, A. R., 348, 422
 Glaser, Maynard, 423
 Goddard, Henry H., 33
 Goheen, Howard W., 55, 162
 Goldenweiser, Alexander, 9, 10
 Good, Carter V., 15, 21, 127
 Goodenough, Florence L., 33, 55, 162, 191
 Goossen, Carl V., 422
 Gordon, Hans C., 22, 406
 Gordon, W. E., 195
 Grambs, Jean E., 414
 Gray, J. Stanley, 14
 Green, Bert F., Jr., 206
 Greenberg, Jacob, 183, 188
 Greene, H. A., 176, 177
 Gregory, C. A., 171, 173
 Grimes, James W., 144
 Grossmickle, Foster E., 126
 Gunler, Walter Scribner, 330, 331
 Guilford, J. P., 105, 123, 135, 144
 Gulliksen, Harold, 25, 55, 134, 135

 Hagen, Elizabeth P., 418
 Haggard, Ernest A., 278
 Haggerty, Lida Harmar, 298
 Haggerty, M. E., 382
 Haines, Mullicent, 60
 Hall, G. Stanley, 49, 50, 52, 404
 Hall, Wilbur, 299
 Hangen, 243

 Harap, Henry, 348
 Harlow, Harry F., 327
 Harris, Albert J., 294
 Harry, David P., Jr., 168, 189
 Hart, 49
 Hartley, Henry H., 406
 Hartshorne, Hugh, 47, 48
 Hartung, Maurice, 327
 Hawkes, Herbert E., 150, 156, 165, 180, 185, 186, 197
 Hawkinson, Mabel J., 336
 Healy, William, 35
 Heal, Walter G., 110, 111, 286, 287
 Heim, Alice W., 327
 Heilmann, Robert A., 371, 422
 Heins, H., 288
 Heisenberg, W., 132
 Helmholtz, Hermann, 8, 31
 Henry, Lyle K., 311
 Henry, Nelson B., 191, 206
 Hensley, Iven H., 139, 164, 197
 Herrick, James B., 12
 Hertz, H. R., 11
 Hevner, Kate, 174
 Hildreth, Gertrude H., 54, 58, 130, 288, 333, 337, 338, 345, 366
 Hilgard, Ernest R., 327, 373
 Hilkert, Robert N., 245
 Hill, George E., 175, 407, 408, 409
 Hoglan, 309
 Hollingshead, Augustus B., 366
 Hollingsworth, Leta S., 284
 Holzinger, Karl J., 418
 Hopkins, Earnest Martin, 28
 Hopkins, Mark, 116
 Horn, Alice, 110, 111, 286, 287
 Horst, Paul, 185
 Howerth, I. W., 8
 Hubbard, Henry D., 247
 Hudelson, 43
 Huey, 53
 Hull, Clark L., 54, 351, 352
 Hulten, C. E., 43
 Hunnecutt, Clarence W., 323
 Hurlock, Elizabeth B., 321, 322, 326
 Huxley, Julian, 7
 Hyde, M. F., 360

 Iron, Arthur I., 327
 Irwin, J. O., 57
 Ivins, Lester S., 401

 Jackson, Robert W. B., 125
 Jacobs, Robert, 135, 162, 191, 246
 Jahoda, Marie, 25, 51, 398, 425
 James, William, 15
 Jefferson, Thomas, 347
 Jesus, 347
 Jevons, W. Stanley, 30, 31
 Johnson, Bess E., 311
 Johnson, Franklin W., 39
 Johnson, Palmer O., 25, 421
 Jones, Edward Safford, 200
 Jones, Harold E., 22, 310, 312
 Jones, Lyle, 418
 Jones, Vernon, 47

- Jordan, A M, 162, 191, 245
 Judd, Charles H, 368
- Kandel, Isaac L, 38, 193, 400
 Katz, David, 9
 Kavruck, Samuel, 55, 162
 Kay, Marjorie E, 337
 Kelley, Truman L, 52-54, 56, 115, 116,
 124, 132, 161, 170, 229, 273, 290, 296,
 298
 Kelley, V H, 176
 kelvin, Lord, 7
 Kemble, Edwin C, 25
 kendall, Maurice G, 105
 kepler, Johann, 7
 keys, Noel, 22, 311, 312, 330
 khayyám, Omar, 4
 Kilpatrick, William H, 17, 18, 356
 King, Harold V, 422
 King, Ronald, 7
 Kinney, L B, 165
 Kirkpatrick, James Earl, 310
 Kitch, Loran V, 309
 Kostick, Max M, 29, 152, 206
 Kraepelin, Emil, 30
 Kruglov, Lorraine, 152
 Kuder, G F, 51, 124, 129, 219, 416, 417,
 419, 421
 Kugle, CIO
 Kuhlmann, Frederick, 33, 182, 208, 288
 Kulp, Daniel H, 22, 312
- Laird, Donald A, 49
 Lampson, Edna E, 321
 Langmuir, Charles R, 103, 134
 Langmuir, Irving, 133
 Lawler, Eugene S, 249
 Learned, William S, 350
 Leary, B E, 141
 Lee, J Murray, 164, 165, 180, 197
 Lee, Richard E, 4
 Lefever, D Welty, 371
 Lehman, Harvey C, 51
 Leiter, Russell G, 422
 Lennon, Roger T, 278, 287
 Lentz, Theodore F, Jr, 47, 154
 Leonard, E A, 240
 Leonard, J Paul, 331, 398
 Leonard, Sterling A, 171
 Lev, Joseph, 105
 Lewin, Lillie, 246
 Ligon, Ernest M, 228, 230
 Lincoln, Abraham, 375, 415
 Lindquist E F, 22, 25, 55, 58, 105, 119,
 121, 127, 135, 139, 142, 150, 154, 156,
 160, 162, 165, 180, 185, 186, 191, 197,
 206, 245, 274, 295, 300, 327, 345, 346,
 371, 417, 440
 Lindvall, Carl M, 398
 Lindzey, Gardner, 176, 190, 244, 421
 Lockhart, Aileene, 22
 Locy, William A, 8
 Longstaff, H P, 311
 Lord Frederic, 420
 Lorge Irving, 20 135 152, 278
 Lundholm H T, 168 178
- McAllister, Jane E, 22
 McCall, William A, 17, 44, 53, 54, 227,
 279, 295, 361, 362, 377
 McClelland, David C, 327
 McConnell, Max, 28
 McConnell, James, 51
 McCullough, Constance M, 345
 McGeoch, John A, 327
 McGregor, J B, 195
 McKinney, H T, 364
 McNamara, Walter J, 135, 160, 162, 181,
 191
 McNemar, Quinn, 32, 52, 91, 108, 110, ,
 111, 278, 284, 286 287, 429
 Mackenzie, Gordon N, 366
 MacKinnon, Donald W, 327
 Madison, James, 402
 Maller, Julius Bernard, 360
 Malthus, Thomas R, 10
 Mann C R, 150, 156, 165, 180, 185, 186,
 197
 Mann, Horace, 29, 38, 45
 Mann, William A, 345
 Marconi, Guglielmo, 11
 Marston, William M, 49
 Martin, Abe, 19
 Massey, Frank J, Jr, 105
 Mather, Kirtley F, 6
 Mathews, C O, 49
 Maxfield, Francis N, 57, 58
 Maxwell, James Clerk, 11
 Mav, Mark A, 47, 48
 Mayer, Joseph R, 9
 Melby, Ernest O, 365
 Mendel, Gregor J, 8
 Menninger, Karl A, 299
 Merrill, Maud A, 34, 282, 288-290, 351
 Messenger, Helen R, 304, 406
 Meumann, Ernst, 31
 Meyer, George, 324
 Meyer, Max, 39
 Miller, W S, 287
 Mills, Lewis H, 168
 Miner, James B, 51
 Mitchell, Claude, 319
 Modley, Rudolph, 273
 Mollenkopf, William G, 371
 Monroe, Walter S, 43, 44, 49, 56, 58, 59,
 113, 114, 165, 193, 200, 297, 325
 Moody, Caesar B, 59
 Moore, Bruce V, 51, 54
 Morgenstern, Oskar, 423
 Morrisett, L N, 407
 Mort, Paul R, 394-396
 Mosser, Charles I, 180
 Mowrer, O Hobart, 327
 Muller, Johannes, 8
 Murphy, Gardner, 348
 Myers, George E, 368
 Myers, M Claire, 180
- Neale, M G, 404
 Nelson M J, 183
 Nemzek, Claude L, 284
 Neumann, John Von, 423
 Newcomb, Theodore M, 327

- Newens, Lyndall Fisher, 196
 Newland, T. Ernest, 337
 Newman, Sidney H., 311
 Nixon, Belle M., 29, 152, 206
 Noll, Victor H., 311
 Norton, John K., 249
 Norvell, Lee, 321
 Nowlis, Vincent, 327

 Odell, C. W., 59, 105, 135, 162, 191, 206
 Oden, Melita H., 366
 Ogburn, William F., 9, 10
 Olson, Helen F., 197
 Omwake, K. T., 178
 Orata, Pedro D., 376
 Orleans, Jacob S., 361
 Osler, Sir William, 338
 Otis, Arthur S., 36, 110, 111, 220, 221, 286, 287, 319
 Otto, Henry J., 348, 358, 359, 364, 365
 Oxtoby, Toby, 372

 Pace, C. Robert, 374, 377, 398, 399
 Panlasiguit, Isidoro, 315
 Parten, Mildred B., 52
 Partridge, E. DeAlton, 360
 Paterson, Donald G., 35, 37
 Payne, William H., 16
 Pearson, Karl, 8, 31, 33, 48, 85, 86, 102, 104, 109, 433, 435
 Pease, Glenn R., 313, 314
 Peattie, Donald Culross, 9
 Peel, E. A., 327
 Peters, Charles C., 243
 Peterson, Joseph, 59
 Phillips, Alexander J., 156
 Pieper, C. J., 168, 171, 177
 Pintner, Rudolph, 35, 37, 44, 59, 287
 Planck, Max, 7, 12, 26, 132
 Plato, 9, 14, 351
 Poffenberger, Albert T., 299
 Postman, Leo J., 327
 Pressey, Sidney L., 49, 129, 312, 327, 340, 363, 364, 366
 Price, Helen G., 180
 Pythagoras, 9

 Quetelet, 10

 Rath, Louis E., 141, 142
 Reavis, W. C., 398
 Reeder, Ward G., 404
 Reeve, Ethel B., 183
 Remmers, Hermann H., 49, 116, 206, 312, 313
 Rice, Joseph M., 20, 21, 38, 45, 57, 59
 Richardson, Helen M., 360
 Richardson, Marion W., 124, 219, 416
 Ricks, James H., Jr., 300
 Riley, John L., 45, 46, 405
 Runland, Henry Daniel, 165, 168, 171, 173, 202
 Roens, Bert A., 371, 415
 Rogers, Carl R., 370, 371
 Rorschach, Hermann, 421
 Roseborough, Mary E., 421

 Ross, C. C., 133, 240, 306, 311, 317, 319, 320, 363
 Rothney, John W. M., 371, 372, 415, 422
 Rousseau, Jean Jacques, 347
 Ruby, Lionel, 26
 Ruch, G. M., 54, 59, 156, 163-165, 170, 193, 195, 294
 Rugg, Harold O., 17, 59
 Rulon, Philip J., 111, 134, 278, 300, 417, 419, 420
 Russell, Bertrand, 5, 6, 11, 12, 19, 26
 Russell, David H., 327, 373
 Ryan, T. M., 182
 Ryan, W. Carson, Jr., 194
 Ryans, David G., 31, 246

 Sacks, Elinor L., 133
 Sarton, George, 26
 Saucier, W. A., 24
 Saupe, Joe L., 398
 Sauvain, Walter Howard, 360
 Scates, Douglas E., 15, 21, 24, 60, 127, 133, 249, 264, 296, 338
 Schrader, William B., 103, 371
 Schrammel, H. E., 182
 Schutte, T. H., 22
 Schwab, Joseph J., 135
 Schwiering, O. C., 345
 Scott, Walter D., 48
 Scott, William O. N., 415
 Scruggs, Sherman D., 330
 Seashore, Carl, 37, 54
 Seashore, Harold G., 134, 245, 300, 371, 419
 Segel, David, 164, 197, 294
 Seguin, Edouard, 35
 Sells, Saul B., 398
 Selover, Margaret, 135, 162, 191, 246
 Seyfert, Warren C., 415
 Shaffer, Laurance F., 120
 Shakespeare, William, 121
 Sharp, George, 421
 Shaw, Duane C., 38
 Sherman, N. H., 174
 Shih, Hu, 4
 Shore, 309
 Siceloff, L. P., 168, 177
 Simpson, Robert G., 346
 Sims, Verner M., 196, 204, 374
 Singer, Arthur, 421
 Smart, Harold R., 5, 14
 Smeltzer, C. H., 312
 Smith, B. Othanel, 5, 11, 48
 Smith, Eugene R., 117, 140, 141, 374, 415, 422
 Sones, W. W. D., 168, 189
 Spanney, Emma, 177
 Spaulding, Geraldine, 183, 188, 246
 Spear, Mary Eleanor, 249, 253, 272, 273
 Spearman, Charles E., 8, 31, 52, 54, 99, 123, 124, 435
 Spence, Ralph B., 209, 210
 Stalnaker, John M., 135, 139, 192, 194, 195, 203-206
 Stanley, Julian C., 55, 118, 124, 162, 206, 244, 284, 300, 371, 417, 420, 422, 452

- Starch, Daniel, 40, 41, 53
 Starkey, Mary L., 246
 Stauffer, Russell G., 346
 Stenquist, John L., 37, 210, 237, 238, 246, 278
 Stephenson, William, 52, 55, 191, 216
 Stern, Wilhelm, 31
 Sternberg, Jack J., 419
 Stevens, S. Smith, 26
 Stoddard, George D., 54, 165, 225, 287
 Stone, C. W., 39, 168, 331
 Stouffer, Samuel A., 24
 Strang, Ruth M., 345, 372
 Strayer, George D., 395
 Strong, E. K., Jr., 51
 Strunk, Mildred, 52
 Stuit, Dewey B., 346
 Stutsman, Rachel, 289
 Suelzt, Ben A., 327
 Super, Donald E., 55, 246, 372
 Sutherland, A. A., 348
 Sykes, Gresham M., 415
 Symonds, Percival, 48, 54, 165
 Swan, J. N., 276

 Taba, Hilda, 305
 Tallmadge, Margaret, 323
 Terman, Lewis M., 30-34, 53, 108, 110, 111, 170, 232, 279, 280, 282-284, 286-288, 290, 299, 345, 351, 363, 366
 Terry, Paul W., 323, 324
 Thelen, Herbert A., 414, 421
 Thisted, M. N., 313
 Thompson, George G., 323
 Thompson, Loring M., 273
 Thorndike, Edward L., 17, 30, 33, 34, 38, 39, 43, 51-53, 113, 121, 135, 237, 260, 279, 295, 323, 376, 400
 Thorndike, Robert L., 116, 284, 300, 400, 415
 Thurstone, Louis L., 37, 38, 49, 52, 59, 144, 221, 281, 289, 418
 Thurstone, Thelma Gwinn, 289, 418
 Tiedeman, David V., 300, 371, 398, 418, 419
 Tiegs, Ernest W., 175, 358
 Tilton, John W., 23
 Tinkelman, Sherman, 152
 Tippet, L. H. C., 105
 Todhunter, Isaac, 19
 Torgerson, Warren S., 206
 Townsend, Agatha, 135, 162, 191, 246, 372
 Travers, Robert M. W., 135, 162, 191
 Travler, Arthur E., 103, 127-129, 135, 162, 183, 191, 195, 211, 243, 246, 300, 334, 341, 344, 345, 372, 378, 398, 415
 Trow, William Clark, 414
 Troyer, Maurice E., 374, 399
 Tsao, Fei, 298
 Tucker, A. C., 240
 Turney, Austin H., 22, 311, 312, 358, 360, 362
 Turrell, Archie M., 371
 Tyler, F. T., 314
 Tyler, Ralph W., 21, 114, 115, 117, 140, 141, 327, 338, 346, 374, 401, 415, 422
 Tyson, Robert, 9

 Updegraff, Ruth, 225
 Upton, Clifford B., 119

 Vaughan, Kenneth W., 245
 Vernon, M. D., 273
 Vernon, Philip E., 135, 176, 190, 221, 244, 421
 Voelker, 47
 Votaw, David F., Sr., 154

 Waldrop, Robert S., 417
 Walker, Helen M., 10, 60, 85, 90, 105, 378, 456
 Warner, Charles Dudley, 37
 Washburne, John Noble, 249
 Washington, George, 20
 Watson, Goodwin, 48, 49, 422, 423
 Watts, Winfred, 406
 Weber, E. H., 8, 31
 Wechsler, David, 55, 215, 279, 282, 289, 418, 422
 Weidemann, Charles C., 194, 196, 198, 200, 203, 204
 Weitzel, Henry I., 371
 Weitzman, Ellis, 135, 160, 162, 181, 191
 Wesley, E. B., 182
 Wesman, Alexander G., 101, 103, 134, 245, 371, 419
 Westaway, F. W., 7, 12, 13, 132
 Wheeler, L. R., 277
 White, Clyde W., 306
 White, Emerson E., 29, 30
 White, Hubert B., 314
 Whitehead, Alfred North, 4, 8, 9
 Whitnev, Frederick L., 26, 399
 Wilbur, Ray Lyman, 338
 Wilder, M., 311
 Wilds, Elmer Harrison, 362
 Wilkins, Laroy W., 363
 Willey, Roy D., 372
 Williams, Donald, 26
 Williams, L. A., 359
 Williamson, E. G., 372
 Willis, Mary, 422
 Wilson, Guy M., 134, 332
 Wilson, Kenneth, 371
 Wissler, Clark, 33
 Witt, Paul, 51, 363, 366
 Wolfe, Dael, 372
 Wood, Ben D., 54, 276, 350
 Woodworth, Robert S., 49
 Woody, Thomas, 356
 Worcester, D. A., 201, 246
 Wright, W. H. E., 174
 Wrightstone, J. Wayne, 117, 184, 187, 200
 Wrinkle, William L., 410, 411, 415
 Wriston, Henry M., 376
 Wundt, Wilhelm, 30, 33
 Wyndham, Harold S., 358, 359

 Young, Kimball, 59
 Yule, G. Udny, 105

 Ziegfeld, Edwin, 377
 Zook, George F., 382

Subject Index

- Abilities of Man*, Spearman, 54
- Ability, defined, 276
- Ability groups, 357-65
- acceleration, 363-65
 - arguments for and against, 358
 - continuous promotion, 365
 - experimental evidence, 358-60
 - retardation, 363-65
 - technique of grouping, 361-63
- Abnormal psychology, France and, 31-34
- Acceleration, 363-65
- Accomplishment quotient (AQ), 298
- Achievement
- intelligence and, comparing, 296-99
 - scores, intelligence and, combining, 298-99
- Achievement tests
- defined, 23
 - diagnostic, development, 44
 - history, 35-46
 - improved examinations, 45-46
 - progress before 1918, 38-39
 - progress since 1918, 43-45
 - unreliability of school marks and examinations, studies in, 39-43
 - intelligence vs aptitude, 418-19
 - norms, use in interpreting scores on, 290-96
 - objective type, development, 44
 - practice, development, 44
 - specific type, development, 44
 - standardized and nonstandardized, advantages and limitations, 216-17
 - survey type, development, 44
 - validation, 111-21
 - criticisms, 113-14
 - curricular versus statistical, 111-13
 - direct vs indirect methods, 115-17
 - item analysis, 117-19
 - standard tests, judging, 119-21
 - Tyler's suggestions, 114-15
- Activity movement, 356-57
- Administration
- ease of, usability, 127-28
 - procedure for, 229-30
 - school, measurement in, function, 21-22
 - test, 225-30
 - time for, 225-27
 - who should administer, 227
- Administrators, school, records and reports for, 240-43
- Age
- chronological, see Chronological age
 - educational, see Educational age
 - increase, criterion of intelligence, 108
 - mental, see Mental age
 - promotion, see Promotion age
 - subject age, see Subject age
- Agencies of public information, ordinary, 402-04
- newspapers, local, 402-03
 - student publications, 403-04
- Allport-Vernon-Lindzey Study of Values, 176, 190, 421
- Alternative-response tests, 174-79
- advantages, 174-75
 - construction, rules and suggestions for, 178-79
 - definition, 174
 - illustrations, 175-78
 - limitations, 174-75
- America, applied psychology and, 33-34
- American Council on Education Psychological Examination, 326
- American Journal of Psychology*, 52
- American Psychological Association, 19, 36, 416
- Annual reports, public relations, 404
- Answer keys, preparing, 158
- Application, ease of, usability, 129-30
- Applied psychology, America and, 33-34
- Applied sciences, measurement in, 11-12
- Appraising Vocational Fitness*, Super, 53
- Aptitude Testing*, Hull, 54
- Aptitude tests
- achievement vs intelligence, 418-19
 - special, 37-38
- AQ, see Accomplishment quotient
- Army Alpha tests, 16, 36-37, 108
- Army Beta tests, 37
- Arthur Adaptation of Leiter International Performance Scale, 422
- Astronomy, measurement in, 6
- Ayers educational index, 115
- Bar graphs, 262, 263, 264
- Barrett-Ryan Literature Test *Silas Warner*, 182

- Bible testing device in 27
Bibliography of Mental Tests and Rating Scales, Hildreth, 54
 Biological sciences measurement in, 7-9
 Books, important 33 55
- CA see Chronological age
 California Achievement Tests, Advanced Battery, 175-76 254-55 256
 California Short-Form Test of Mental Maturity, 110, 287
 Capacity, defined, 276-77
 Central tendency, concept of test data, 69, 73-74
 measure 75
 Character Education Inquiry, 47
 Character measurement, history 46-52
 beginnings, 46
 development, 46-48, 51-52
 interview, 51
 questionnaire, 49-51
 rating scales, 48
 Children's Apperception Test (CAT), 421
 Chinese civilization, testing devices in, 27-28
 Chronological age (CA), 279-80
 Circle graphs, 263
 Clapp-Young self marking tests 129
 Classification
 ability groups, 357-65
 acceleration, 363-65
 arguments for and against, 358
 continuous promotion, 365
 experimental evidence 358-60
 individual and group instruction, 357
 retardation 363-65
 technique of grouping, 361-63
 activity movement, 356-57
 human variability, 347-56
 group differences, 348-50
 individual differences, 351-52, 353-56
 problem, 347-48
 trait variability, 352-53
 test results, 61-69
 rank order, 61-62
 Class interval, selecting, frequency distribution 64
 Coefficient of correlation, see Product-moment coefficient of correlation
 Coherency, criterion of intelligence 109
 Cole-von Borgersrode Scale for Rating Standardized Tests, 222-24
 College Board Review 192
 College Entrance Board Examinations, 192 195 204 305
 Colorado experiment, reporting to parents 410 11
 Column diagram 258-59
 Committee on Standards for Graphic Presentation 272
 Completion tests 170-74
 advantages 170
 construction rules and suggestions for 172 74
 definition 170
 Completion tests (Cont.)
 illustrations, 170-71
 limitations, 170
 Computations, concepts versus, quantitative data, 69-75
 Concentration objective of instruction, 145
 Concepts
 computations versus, quantitative data, 69-75
 co-relationship or concomitant variation 85-90
 Concomitant variation, concept 85-90
 Concurrent validity, 417
 Confidence, stage in use of tests, 57-58
 Congruent validity, 417
Construction and Analysis of Achievement Tests, Adkin et al., 55
 Construction of tests, see Test construction
 Content validity, 417
 Continuous promotion, 365
 Co-operation, objective of instruction, 146
 Cooperative Achievement Tests, 112, 140, 291
 Cooperative English Test, 171, 422
 Cooperative French Test Junior Form 1946, 183, 188-89
 Cooperative Plane Geometry Test, Revised Series Q, 177
 Cooperative Solid Geometry Tests, 178
 Cooperative Study of Secondary School Standards, 374-75, 376, 378-80, 383, 384, 388, 390, 394, 395
 Cooperative Test of Social Studies Abilities 182, 184, 187-88
 Co-relationship, concept, 85-90
 Correlation, 85-90
 coefficient, see Product-moment coefficient of correlation
 rank, 95-98
 Cost usability of measuring instrument, 130
 Creativeness, objective of instruction, 144
 Criterion problem, 417-18
 Critical caution, stage in use of tests, 58
 Crude scores, see Raw scores
 Curiosity, stage in use of tests 57
 Curricular validity, 101
 statistical versus, 111-13
- Davis Eells Games for Grades I-VI, 422
 Deciles, 75
 Decision making, information and, 423-24
 Derived scores, 288-90
 intelligence quotient, see Intelligence quotient
 raw scores versus, 278-79
 Deviation
 quartile, see Quartile deviation
 standard see Standard deviation
Diagnosing Personality and Conduct Syndromes, 48
 Diagnosis, 328 45
 causes of errors, locating 337-41
 individuals needing locating, 332 33
 levels 332

- Diagnosis (*Cont*)
 nature of difficulty, locating 333-37
 nature of educational, 328-30
 preventive, 315
 problem, 328-31
 remedial procedures, 341-45
 techniques, 332-45
 value in education, 330-31
- Differential Aptitude Tests, 418-19
- Difficult prediction, 419
- Difficulty, measure, item-analysis, 440
- Discrimination, measure, item-analysis
 437-40
- Discrimination table, item-analysis, 447-51
- Distribution, frequency, *see* Frequency distribution
- Draft, preliminary, preparing the test, 147-48, 149
- EA, *see* Educational age
- Education
 meaning of, 13-16
 measurement in, 13-25
 achievement tests history, 38-46
 character, personality and interest tests, history, 46-52
 function, 21-22
 historical development, 27-38
 intelligence tests, history, 30-38
 place of, 16-18
 publications, important 52 55
 recent tendencies, 56-58
 types, 23-25
 views of, 17
 reputation in, 19-20
 research in, 19, 20
 rhetoric in, 19
 three Rs in, 18-21
- Educational, *Psychological, and Personality Tests of 1933-35*, Buros, 54
- Educational Administration and Supervision, 53
- Educational age (EA)
 educational quotient versus, 290-91
 limitations, 291-92
 uses, 291
- Educational and Psychological Measurement, 53
- Educational Guidance, Kelley 53
- Educational Measurement, Lundquist 22, 55
- Educational Measurement, Starch 53
- Educational program, evaluating 384-94
- Educational quotient (EQ)
 educational age versus 290-91
 limitations, 292-93
 use, 292-93
- Educational Records Bureau 243
- Educational Test Bureau 257
- Eight-Year Study of the Progressive Education Association, 117 140 374
- Elementary schools, principles of evaluation for, 380-81
- England statistical methods and 31
- I Q *see* Educational quotient
- Error
 causes of, diagnosis, 337-41
 measures of, 101-03
 measurement, 101, 102-03
 sampling, 101, 103
 technique, 101, 102
- Essay examinations, 192-205
 advantages, 196-97, 200-01
 grading by sorting 204-05
 improving suggestions for, 197-205
 advantages, special 200-01
 construction, 198-99
 scoring 202-04
 use, 198-99
- Institutions 193-95
 preparing students to take 201-02
 reliability, 193-95, 196
 usability, 195, 196
 validity, 193 196-97
- Essentials of Psychological Testing*, Cronbach, 55
- Evaluation
 in guidance, 367-71
 co-operative venture, 370-71
 importance 367-68
 meaning 367-68
 measurement in, 369-70
- of schools, 373-98
 Cooperative Study of Secondary School Standards, 378-80
 difficulty, 376-77
 educational program, 384-94
 importance 375-76
 index of variation, 397-98
 measurement and 373-75
 organization and plant 395-97
 philosophy of school 383 84
 principles of general 380-83
 problem 373 80
 teaching efficiency, 377-78
 tests in use 379
- qualitative technique 420-22
 semi-quantitative technique, 420-22
 test, 159-62
- Every Pupil Test in Physics, 187
- Examinations
 essay, *see* Essay examinations
 school
 improved 45-46
 unreliability, studies in 39-43
- Exhibits school public information 413
- Expectancy tables 100-01
- Experimental psychology German and 30-31
- Factor analysis 418
- Fourth Mental Measurements Yearbook* (1953) Buros 55 221
 classification of tests in 218-219
- Forty-Fifth Yearbook of the National Society for the Study of Education* 422
- France abnormal psychology and 31-34
- Frequency distribution or table 62-69
 form 66-67
 graphical representation 258-64
 histograms 262 263 264

Frequency distribution or table (*Cont.*).

graphical representation (*Cont.*)

circle graphs, 263

column diagram, 258-59

frequency polygon, 259-60

histogram, 258-59

pictographs, 263

pie graphs, 263

skewed curves, 262-63

smooth curve, 260-62

symmetrical curves, 262-63

typewriter graphs, 263

which graph is best?, 263-64

making, 64-66

scattergram or scatter diagram, 67-69

two-way, 67-69

Frequency polygon, 259-60

Frequency table, *see* Frequency distribution

George Washington University English Literature Test, 178

Germany, experimental psychology and, 30-31

Gestalt school of psychology, 9, 16-17

Grade norms

limitations, 293-94

use, 293-94

Grading, *see* Scoring

Graphs, 247-73

constructing, suggestions for, 271-73

distributions, two or more representing, 264-71

central tendencies and variabilities of, 269-71

central tendencies of, 267-69

entire distributions, 264

frequency, *see under* Frequency distribution

percentile curves, use, 265-66

polygons, use, 264-65

frequency distribution, representing, 258-64

bar graphs, 262, 263, 264

circle graphs, 263

column diagram, 258-59

histogram, 258-59

pictographs, 263

pie graphs, 263

polygon, 259-60

skewed curves, 262-63

smooth curve, 260-62

symmetrical curves, 262-63

typewriter graphs, 263

which graph is best?, 263-64

record of an individual, representing, 254-58

profiles for series of subjects, 254-58

profiles of single subject, 254

value, 247-54

attention getting, 247-49

points clarified, 249

retention aided, 249-54

Gregory Tests in American History, 171

Group

ability *see* Ability groups

Group (*Cont.*)

differences, 348-50

instruction, 357

Grouped frequency distribution, 64, 65

Guidance, evaluation in, 367-71

co-operative venture, 370-71

importance, 367-68

meaning, 367-68

measurement in, place of, 369-70

Hemis Mental Growth Units, 288

Higher institutions, principles of evaluation for, 352-83

Histogram, 258-59

Holzing-Crowder Uni-Factor Tests, 418

Homogeneous groups, *see* Ability groups

How to Measure in Education, McCall, 54

Hudelson Scale, 43

Human variability

educational significance, 347-56

group differences, 348-50

individual differences, 351-52

educational provisions for, 353-56

nature, 347-56

problem, 347-48

trait variability, 352-53

IBM General Purpose Answer Sheet, 159

Improvement of the Written Examination, Ruch, 54

Index of variation, evaluation of schools, 397-98

Individual

differences, 351-52

educational provisions for, 353-56

instruction, 357

profile chart, 239

Information, decision-making and, 423-24

Initiative, objective of instruction, 144-45

Instruction

individual and group, 357

measurement in, 303-425

classification and promotion, 347-65

diagnosis, 328-45

evaluation in guidance, 367-71

evaluation of schools, 373-98

function, 21-22

motivation and practice in testing, 303-27

public relations, 400-15

trends, present, 416-24

Objectives, categories, 141-42

concentration, 145

co-operation, 146

creativity, 144

initiative, 144-45

interest, 145

judgment, 145-46

motivation, 145

outcomes, provision for evaluating, 141-46

relation of measurement to motivation in, 304-06

teaching efficiency, evaluating, 377-78

Intelligence

achievement and, comparing 296-99

- Intelligence (*Cont*):
 meaning, 108
 scores, achievement and, combining, 298-99
 Terman criteria, 108-09
- Intelligence quotient (IQ)
 advantages, 284-85
 computation, 281-83
 equating values, table for, 286
 interpretation, 283-84
 limitations, 285-88
 mental age vs., 279-80
- Intelligence tests
 achievement vs. aptitude, 418-19
 defined, 23
 history, 30-38
 abnormal psychology, France and, 31-34
 applied psychology, America and, 31-34
 children of necessity, 34-37
 experimental psychology, Germany and, 30-31
 special aptitude tests, 37-38
 statistical methods, England and, 31
 norms, use in interpreting scores on, 279-90
 validation, 108-11
 individual vs. group, 109-11
 meaning of intelligence, 108
 Terman criteria, 108-09
- Interest, objective of instruction, 145
- Interest measurement, history, 46-52
 beginnings, 46
 development, 46-48, 51-52
 interview, 51
 questionnaire, 49-51
 rating scales, 48
- International Institute of Teachers College, Columbia University, 194
- Interpretation, ease of, usability, 129-30
- Interpretation of Educational Measurements*, Kelley, 54
- Interviews, character, personality and interest measurement, 51
- Introduction to the Theory of Mental and Social Measurements*, Thorndike, 53
- Iowa Academic Testing Program, 304
- Iowa Every-Pupil Tests in Basic Skills, 177
- Iowa Placement Examinations, 37, 168
- Iowa Silent Reading Tests, New Edition, 176, 254, 255, 263
- IQ, *see* Intelligence quotient
- Item-analysis procedure, simplified, 436-53
 difficulty, measure, 440
 discrimination, measure, 437-40
 discrimination table, 447-51
 illustrative analysis, 440-47
 items, preparing, 436-37
 mean obtaining, 452
 reliability coefficient obtaining, 452
 standard deviation obtaining, 452
 validation, 117-19
- Items, test
 analysis, *see* Item-analysis procedure
 preparing, 436-37
 ranking, 454-55
- Journal of Applied Psychology*, 53
Journal of Consulting Psychology, 53
Journal of Educational Psychology, 52
Journal of Educational Research, 53
Journal of Genetic Psychology, 52
- Judgment, objective of instruction, 145-46
- KR No. 20, 124
- Kuder and Richardson formulas, 124
- Kuder Preference Record, 51, 129, 421
- Kuhlmann-Finch Intelligence Tests, 182
- L'Année Psychologique*, 31, 52
- Learning
 amount and quality, relation of measurement to, 309-23
 awareness of final examination, 312-15
 frequency of tests, 309-12
 knowledge of results combined with other incentives, 319-23
 knowledge of test scores, 315-19
 motivation in, relation of measurement to, 306-25
 type of, relation of measurement to, 323-25
- Leiter International Performance Scale, Arthur Adaptation, 422
- Letters to parents, 410-13
 Colorado experiment, 410-11
 suggestions for, 410
 University of Chicago High School System, 411-13
- Limitations
 alternative-response tests, 174-75
 completion tests, 170
 educational age, 291-92
 educational quotient, 292-93
 essay examinations, 193-95
 grade norms, 293-94
 intelligence quotient, 285-88
 matching tests, 186
 measurement, 12-13
 mental age, 280-81
 multiple-choice tests, 180-81
 sample-recall tests, 167
- Limits of classes determining, frequency distribution, 64
- Local norms, value, 295-96, 297
- MA, *see* Mental age
- Marks, school unreliability, studies in, 39-43
- Matching tests, 186-90
 advantages, 186
 construction rules and suggestions, 189-90
 definition, 186
 illustrations, 187-89
 limitations, 186

- Mean, 73
 computing, from scattergram data, 94
 finding, 78-81
 median versus, 74-75
 obtaining, item-analysis, 452
 short way to compute, 80
- Measurement
 errors in, controlling, 13
 errors of, measures, 101, 102-03
 importance, 3-4
 in applied sciences, 11-12
 in biological sciences, 7-9
 in education, 13-25
 achievement tests, history, 38-46
 character, personality and interest tests, history, 46-52
 function, 21-22
 historical development, 27-58
 intelligence tests, history, 30-38
 place of, 16-18
 publications, important, 52-55
 recent tendencies, 56-58
 types, 23-25
 views of, 17
 in guidance, 369-70
 in instruction, 303-425
 classification and promotion, 347-65
 diagnosis, 328-45
 evaluation in guidance, 367-71
 evaluation of schools, 373-98
 motivation and practice in testing, 303-27
 public relations, 400-15
 trends, present, 416-24
 in modern world, 3-23
 in physical sciences, 6-7
 in science, 4-13
 in social sciences, 9-11
 limitations, 12-13
 problem of, 3-134
 generalizations regarding, 131-34
 relation to amount and quality of learning, 309-23
 awareness of final examination, 312-15
 frequency of tests, 309-12
 knowledge of results combined with other incentives, 319-23
 knowledge of test scores, 315-19
 relation to motivation, 304
 in learning, 306-25
 in teaching, 304-06
 relation to type of learning, 323-25
 teacher and, purpose, 304
 teaching emphasis and, 304-06
- Measurement in Higher Education*, Wood, 54
- Measurement in Secondary Education*, Symonds, 54
- Measurement of Adult Intelligence*, Wechsler, 55
- Measurement of Intelligence*, Terman, 34, 53
- Measurement of Intelligence*, Thorndike, 54
- Measuring instrument, satisfactory, characteristics, 106-34
- importance of problem, 106
- reliability, 121-27
 determining, methods, 122-24
 importance, 121-22
 interpretation, 124-25
 meaning, 121
 objectivity and, 125-27
- usability, 127-31
 administration, case, 127-28
 application, case, 129-30
 cost, 130
 interpretation, case, 129-30
 meaning, 127
 mechanical make-up, 130-31
 scoring, case, 128-29
- validity, 107-21
 achievement tests, 111-21
 considerations, general, 107-08
 intelligence tests, 108-11
 meaning, 107
- Mechanical make-up of tests, usability, 130-31
- Median, 73
 determining, from scattergram data, 94-95
 finding, 75-78
 mean versus, 74-75
 process of locating, 76
- Mental age (MA)
 concept, 32
 advantages, 280
 intelligence quotient versus, 279-80
 limitations, 280-81
- Mental Measurements Yearbooks*, 54, 120, 121, 218
- Mental quotient, 31
- Mental Testing*, Goodenough, 55
- Methodology of Educational Research*, Good et al, 127
- Metropolitan and Standard Achievement Tests, 291
- Midscore, 75-76
- Mind*, 52
- Mode, 73
 finding, 75
- Modern School Achievement Tests, Language Usage, 182
- Motivation
 experiment in, 306-07
 limitations on, 307-08
 types, 308-25
 importance, 303
 meaning, 303-04
 objective of instruction, 145
 practice effect, 326-27
 problem, 303-04
 relation of measurement to, 304
 in learning, 306-23
 in teaching, 304-06
 studies, educational implications, 325-26
 for educational practice, 325-26
 for educational theory, 325

- Multiple-choice tests, 179-80
 - construction rules and suggestions for, 181-86
 - definition, 179-80
 - illustrations, 181-84
 - limitations, 180-81
 - possibilities, 180-81
- National Council of Teachers of English, 119
- National Education Association, 374, 380
- National norms, 295
- Natural sciences, measurement in, 8-9
- Nelson High School English Test, 181-82, 183
- Newspapers, local, agency of public information, 402-03
- New York State Regents, 305
- Nineteen Forty Mental Measurements Yearbook*, 55
- Nonstandardized tests, standardized vs., 274-75
- Normal curve, 260
- Normal Progress Chart, 257-58
- Norms
 - grade, 293-94
 - intelligence and achievement, comparing, 296-99
 - interpreting scores on achievement tests, use, 290-96
 - interpreting scores on intelligence tests, use, 279-90
 - interpreting scores on personality tests, use, 299-300
 - local, value, 295-96, 297
 - national, 295
 - percentile, 294-95
 - scores, raw and derived, 276-79
 - standards and, 274-76
- North Central Association of Colleges and Secondary Schools, 382
- Northwestern Intelligence Tests, 422
- Novel tests and items, 422-23
- Objective tests
 - alternative-response, 174-79
 - advantages, 174-75
 - construction, rules and suggestions for, 178-79
 - definition, 174
 - illustrations, 175-78
 - limitations, 174-75
 - completion, 170-74
 - advantages, 170
 - construction, rules and suggestions for, 172-74
 - definition, 170
 - illustrations, 170-71
 - limitations, 170
 - construction, principles, 163-91
 - frequency of use by teachers, 163-64
 - matching, 186-90
 - advantages, 186
 - construction rules and suggestions, 189-90
 - definition, 186
 - Objective tests (*Cont*)
 - matching (*Cont*)
 - illustrations, 187-89
 - limitations, 186
 - multiple-choice 179-86
 - construction, rules and suggestions for, 184-86
 - definition, 179-80
 - illustrations, 181-84
 - limitations, 180-81
 - possibilities, 180-81
 - rearrangement, 190-91
 - sample-recall, 167-70
 - advantages, 167
 - construction, rules and suggestions for, 169-70
 - definition, 167
 - illustrations, 167-69
 - limitations, 167
 - types, 163
 - validity and reliability, comparative, 164-67
 - Objectivity, reliability and, 125-27
 - Occupations*, 53
 - Official publications, public relations, 401-06
 - annual reports, 404
 - special reports, 405-06
 - Ogive, 260
 - Ohio State University Psychological Test, 129
 - Opinion, public, mobilizing, 414-15
 - Organization, school, evaluating, 395-97
 - Otis Quick-Scoring Mental Ability Test, 287
 - Otis Scales for Rating Standard Tests, 220
 - Otis Self-Administered Test of Mental Ability, 287
 - Otis Self-Administering Higher Examination, 110-11
 - Parents
 - letters to, 410-13
 - opinion of, sampling, 414-15
 - reports to, 245
 - Parent-teacher association, public information, 413-14
 - Parent-Teacher Association, 245
 - PC, *see* Personal Constant
 - PEA Interpretation of Data Test, 422
 - Pedagogical Seminary*, 52
 - Percentile curve, 260, 265-66
 - Percentile norms, 294-95
 - Percentile rank, 288-89
 - Percentiles
 - computation, 78
 - D, measure of variability, 84-85
 - Personal Constant (PC), 288
 - Personality tests, norms, use in interpreting scores on, 299-300
 - Personality measurement, history, 46-52
 - beginnings, 46
 - development, 46-48, 51-52
 - interview, 51
 - questionnaire, 49-51
 - rating scales, 48

- Personnel and Guidance Journal*, 53
Personnel Selection, Tests and Measurement Techniques, Thorndike, 55
 Philosophy of the school, evaluating, 383-84
 Physical sciences, measurement in, 6-7
 Physics, measurement in, 7
 Pictographs, 263
 Pie graphs, 263
 Pintner General Ability Tests, 287
 Pintner-Paterson Performance Scale, 35, 37
 Planning the test, 140-47
 administration conditions, considering, 147
 emphasis in course, reflecting proportion of, 146-47
 evaluating outcomes of instruction, provision for, 141-46
 purpose to be served, considering, 147
 Plant, school, evaluating, 395-97
 'Platform for the Use of Standard Tests,' 211-12
 Polygons, use, graphical representation, 264-65
 PrA, *see* Promotion age
 Predictive validity, 417
 Preliminary draft, preparing the test, 147-48, 149
 Preparing the test, 147-55
 arranged in ascending order of difficulty, 151-52
 difficulty of items 148-49
 directions to pupil, 153-54
 particular type of items placed together, 151
 pattern of responses avoiding regular sequence in, 152
 phrasing of items, 149-50
 preliminary draft, 147-48
 preliminary draft items, 149
 revision, critical 149
 types of items, 148
 whole content functions in determining answer, 150-51
 written record of responses, provision for, 152-53
 Preventive diagnosis 345
 Primary Mental Ability tests (PMA), 219, 418
Principles of Science, Jevons, 30
 Product-moment coefficient of correlation, r , 86-90
 computing from scattergram, 93-94
 interpreting, 98-100
 magnitude or size, 98-99
 obtaining, 90
 relationship represented by, 89
 reliability, 101
 sign, 98
 validity, 101, 109
 Professional journals 52-53
 Profile chart, individual, 239
 Profiles, 254
 series of subjects, 264-58
 single subject, 264
 warnings concerning 243-45
 Progressive Education Association, 117, 140, 374, 422
 Promotion
 acceleration and retardation, 363-65
 continuous, 365
 Promotion age (PrA), 298
 Promotion quotient (PrQ), 298
 PrQ, *see* Promotion quotient
 Psychograph, 254
 Psychology
 abnormal, *see* Abnormal psychology
 applied, *see* Applied psychology
 experimental, *see* Experimental psychology
 Gestalt school 9, 16-17
 Psychology Colloquium, University of Wisconsin, 423
Psychology of Musical Talent, Seashore, 54
Psychom trials, 53
 Publications
 important, 52-55
 books, 53-55
 professional journals, 52-53
 public relations, agencies of public information
 newspapers, local, 402-03
 official, 404-06
 student publications, 403-04
 Publicity, 401
 Public opinion, mobilizing, 414-15
 Public relations, 400-15
 agencies of public information, ordinary, 402-04
 newspapers, local, 402-03
 student publications 403-04
 letters to parents, 410-13
 Colorado experiment, 410-11
 suggestions for, 410
 University of Chicago High School System, 411-13
 official publications, 404-06
 annual reports, 404
 special reports, 405-06
 parent-teacher association, 413-14
 principal sources, 402
 problem, 400-02
 programs meaning, 401-02
 public opinion, mobilizing, 414-15
 report cards, 406-12
 Hill's study, 407-08
 trends in, 406-07
 school exhibits, 413
 school visitation, 413
 Public School Attainment Tests for High School Entrance, 171
 Publishers of standardized tests, 464-65
 Pupils, report to, 237
 Qualitative evaluation technique, 420-22
 Quantitative data
 concepts versus computations, 69-75
 central tendency, 69, 73-74, 75
 eleven categories, 72
 four categories, 71
 grading dilemma, 69-70
 mean vs median, 74-75

Quantitative data (*Cont*)
 concepts versus computations (*Cont*)
 one category, 70
 three categories, 71
 two categories, 70-71
 variability, 69, 71-73

elementary notions concerning, 69-75

mean, 73

 finding, 78-81

 median vs., 71-75

median, 73

 finding, 75-78

 mean vs., 71-75

mode, 73

 finding, 75

percentiles, computation, 78

Quartile deviation, 72

 variability measure, 81-83

Quartiles, 77

Questionnaires, character, personality, and
 interest measurement, 49-51

Quotients

 accomplishment, *see* Accomplishment
 quotient

 educational, *see* Educational quotient

 intelligence, *see* Intelligence quotient

 mental, *see* Mental quotient

 promotion, *see* Promotion quotient

 subject, *see* Subject quotient

Range, 72

 determining, frequency distribution, 61

 variability measure, 81

Rank correlation, 95-98

Ranking test items, 454-55

Rank order, 61-62

Rating scales, character, personality, and
 interest measurement, 48

Raw scores, derived scores vs., 278-79

Rearrangement tests, 190-91

Records, 236-45

 for administrators, 240-43

 graphical representation, 254-58

 to teachers, 238-40

Relationship, measures 85-101

 coefficient of correlation, r , 86-90

 computing from scattergram, 93-94

 interpreting 98-100

 obtaining, 90

 reliability coefficient, 101

 validity coefficient, 101

 co-relationship or concomitant varia-
 tion, concept, 85-90

 expectancy tables, 100-01

 rank correlation, 95-98

 scattergram, constructing, 90-93

 means from, computing, 94

 medians, determining, 94-95

r from, computing, 93-94

 standard deviations from, computing,
 94

Reliability

 coefficient, 101

 obtaining, item-analysis, 452

Reliability (*Cont*)

 determining, methods, 122-24

 with one test form, 123-24

 with two test forms, 122

 essay examination 193-96, 196

 importance, 121-22

 interpretation of test, 124-25

 meaning, 121

 objective tests, 164-67

 objectivity and, 125-27

 present trend, 416

 quality of satisfactory measuring instru-
 ment, 121-27

Remedial procedures diagnosis, 341-45

Report cards, 406-12

 Colorado experiment, 410-11

 Hill's study, 407-08

 trends in, 406-07

 University of Chicago High School Sys-
 tem, 411-13

Reporting to parents, *see* Letters to par-
 ents

Reports, 236-45

 annual, public relations, 404

 for administrators, 240-43

 special, public relations, 405-06

 to parents or public, 245

 to pupils, 237

 to teachers, 238-40

Reputation, in education 19-20

Research in education, 19, 20

Results, test, *see* Test results

Retardation, 363-65

Retesting 236

Review of Educational Research, 22, 47 in

Revised Stanford-Binet Intelligence Scale,
 55, 215, 282-81

Revision, critical preparing the test, 149

Rhetoric in education, 19

Sampling, errors of, measures, 101, 103

 'Scale Book,' Fisher's, 38

Scaled test, defined, 24

Scales

 rating *see* Rating scales

 test distinguished from, 24

Scatter, *see* Variability

Scatter diagram, *see* Scattergram

Scattergram, 67-69

 constructing, 90-93

 means from computing 94

 medians, determining 94-95

 negative correlation illustrating, 87

r from, computing 93-94

 standard deviations from, computing
 94

School and Society 53

Schools evaluation, 373-98

 Cooperative Study of Secondary School

 Standards 378-80

 difficulty, 376-77

 educational program, 384-94

 importance 375-76

 index of variation, 397-98

 measurement and, 373-75

 organization and plant 395-97

- Schools, evaluation (*Cont*)
 philosophy of, 183-84
 principles of, general, 350-83
 for elementary schools, 350-81
 for higher institutions, 382-83
 for secondary schools, 351-82
 problem, 173-80
 teaching efficiency, 377-78
 tests in, use, 379
- Science, measurement in, 4-13
- Science Research Associates (SRA)
 Non-Verbal test, 110-11
 Primary Mental Ability tests, 110-11, 257
 self-scoring test, 129
- Scientific method, 4-6
- Scores
 analyzing and interpreting, 234-35
 definition, 276-78
 derived, *see* Derived scores
 intelligence and achievement, combining, 298-99
 interpreting on achievement tests, use of norms in, 280-96
 interpreting on intelligence tests, use of norms in, 279-90
 interpreting on personality tests, use of norms in, 299-300
 raw vs derived, 278-79
 sigma, 289-90
 standard, 85, 289-90, 295
 T-scores, 295
 Z-scores, 85, 289-90
- Scoreze, 129
- Scoring the tests, 230-34
 ease of, usability, 128-29
 essay examination, 202-04
 by sorting, 204-05
 procedure, 176-58
 rules preparing, 158
 techniques used, 231-34
 who should score, 230-31
- Scott Man-to-Man Scale, 18
- Seashore Test of Musical Talent, 37
- Secondary schools, principles of evaluation for, 381-82
- Selected References on Test Construction, Mental Test Theory, and Statistics, 1929-49*, Goheen-Kavruck, 55
- Semi-quantitative evaluation technique, 420-22
- Seven Seals of Science*, Mayer, 9-10
- Seventeenth Yearbook of the National Society for the Study of Education*, 53
- Sigma scores, 289-90
- Simple-recall tests, 167-70
 advantages, 167
 construction, rules and suggestions for, 169-70
 definition, 167
 illustrations, 167-69
 limitations, 167
- Skewed curves, 262-63
- Smooth curve, 260-62
- Social sciences, measurement in, 9-11
- Sones-Harry High School Achievement Test, 168, 189
- Spearman-Brown formula, 124
- Special aptitude tests, 37-38
- Special reports and publications, public relations, 405-06
- Specific determiners, 149
- Square roots, computation, 456-58
- Standard deviation, 72
 computing from scattergram data, 94
 obtaining, item-analysis, 452
 practical uses, 85
 simplified way to compute, 84
 variability measure, 83-85
- Standardized tests
 nonstandardized vs., 274-75
 publishers, 464-65
- Standards, norms and, 274-76
- Standard scores, 85, 289-90, 295
- Stanford Achievement Test, 44, 170-71
- Stanford-Binet scale, 33-34, 127
- Statistical analysis, test results, 60-104
 classification and tabulation, 61-69
 frequency table, 62-69
 rank order, 61-62
 considerations, general, 60-61
 error, measures, 101-03
 measurement, 101, 102-03
 sampling, 101, 103
 technique, 101, 102
 quantitative data, 69-75
 concepts versus computations, 69-70
 relationship, measures, 85-101
 coefficient of correlation, 86-90, 93-94, 98-100
 co-relationship or concomitant variation concept, 85-90
 expectancy tables, 100-01
 rank correlation, 95-98
 reliability coefficient, 101
 scattergram data, 90-95
 validity coefficient, 101
 variability or scatter, measures, 81-85
 percentile *D*, 84-85
 quartile deviation, 81-83
 range, 81
 standard deviation, 83-85
- Statistical methods, England and, 31
- Statistical validity, 101
 curricular vs., 111-13
- Statistics
 fifty questions, 429-35
 answers to, 459-63
 importance, 60
 in a capsule, 60-61
- Status validity, 417
- Stenquist Test of General Mechanical Ability, 37
- Stone Reasoning Test in Arithmetic, 39, 168
- Strong Vocational Interest Blank, 51
- Student publications, agencies of public information, 403-04
- Subject age, 291
- Subject quotient, 291
- Symmetrical curves, 262-63

- T-scores, 295
 Tabulation, test results, 62
 frequency table or distribution 62-69
 form, 66-67
 making, 64-66
 scattergram, 67-69
 two-way, 67-69
 Teacher
 measurement and, purpose, 304
 records and reports to 238-40
Teachers College Record, 52
 Teaching, *see* Instruction
 Teaching emphasis measurement and, 304-06
 Technique, errors of, measures, 101, 102
 Terman criteria, intelligence 108-09
 Terman Group Test of Mental Ability, 279
 Terman-McNemar Test of Mental Ability, 110-11, 287
 Test construction, 139-205
 essay examination, 198-99
 evaluating the test, 159-62
 objective tests, principles 163-91
 alternative-response, 178-79
 completion, 172-74
 frequency of use by teachers, 163-64
 matching 189-90
 multiple-choice, 184-86
 simple-recall, 169-70
 types, 163
 validity and reliability, comparative, 164-67
 planning the test, 140-47
 conditions of administration, consideration, 147
 emphasis in course, reflecting proportion of, 146-47
 evaluating outcomes of instruction provision for, 141-46
 purpose to be served consideration, 147
 preparing the test, 147-55
 arranged in ascending order of difficulty, 151-52
 difficulty of items, 148-49
 directions to pupil 153-54
 particular type of items placed together, 151
 pattern of responses, avoiding regular sequence in, 152
 phrasing of items, 149-50
 preliminary draft, 147-48
 preliminary draft items, 149
 revision, critical, 149
 types of items, 148
 whole content functions in determining answer, 150-51
 written record of responses, provision for, 152-53
 principles, general, 139-62
 problem, importance 139-40
 recent tendencies 56-57
 trying out the test, 155-58
 answer keys, 158
 conditions for insuring normal 155
 Test construction (*Cont*)
 trying out the test (*Cont*)
 scoring procedure, 156-58
 scoring rules, 158
 time allowance, 155-56
 Testing program, 209-300
 administering the tests, 225-30
 procedure for, 228-30
 time for 225-27
 who should administer, 227
 considerations general, 209-12
 co operative, 213
 definite, 214
 graphical representation 247-73
 constructing, suggestions for, 271-73
 distributions two or more, representing, 264-71
 frequency distribution, representing 258-64
 record of an individual, representing 254-58
 value 247-54
 norms uses and limitations, 274-300
 intelligence and achievement, comparing 296-99
 interpreting scores on achievement tests, use, 290-96
 interpreting scores on intelligence tests, use, 279-90
 interpreting scores on personality tests use, 299-300
 raw scores and derived scores, 276-79
 standards and norms, 274-76
 plan for elementary school, 210
 practical 213-14
 profiles warnings concerning, 243-45
 purpose of determining, 212-14
 records 236-45
 for administrators, 240-43
 to teachers, 238-40
 reports 236-45
 for administrators 240-43
 to parents or public, 245
 to pupils 237
 to teachers 238-40
 results, applying 235-36
 retesting 236
 scores analyzing and interpreting, 234-35
 scoring the tests, 230-34
 techniques used, 231-34
 who should score, 230-31
 selecting tests, 214-24
 procedure for, 218-24
 type of tests 215
 who shall select, 214-15
 steps in 209-45
Testing School Children, Stepienson, 55
 Test results statistical analysis, 60-104
 applying 235-36
 classification and tabulation, 61-69
 frequency table 62-69
 rank order, 61-62
 considerations general, 60-61

Test results (*Cont*)

- error, measures, 101-03
- measurement, 101, 102-03
- sampling, 101, 103
- technique, 101, 102
- quantitative data, 69-75
 - concepts vs computations, 69-75
 - mean, finding, 78-81
 - median, finding, 75-78
 - mode, finding, 75
 - percentiles, computing, 78
- relationship, measures, 80-101
 - coefficient of correlation, 86-90, 93-94, 98-100
 - co-relationship or concomitant variation, concept, 85-90
 - expectancy tables, 100-01
 - rank correlation, 95-98
 - reliability coefficient, 101
 - scattergram data, 90-95
 - validity coefficient, 101
- variability or scatter, measures, 81-85
 - percentile, *D*, 84-85
 - quartile deviation, 81-83
 - range, 81
 - standard deviation, 83-85

Tests

- achievement, *see* Achievement tests
- administering, 225-30
- alternative-response, *see* Alternative response tests
- aptitude, *see* Aptitude tests
- completion, *see* Completion tests
- construction, *see* Test construction
- essay examinations, *see* Essay examinations
- evaluating, 159-62
- intelligence, *see* Intelligence tests
- matching, *see* Matching tests
- multiple-choice, *see* Multiple-choice tests
- nonstandardized, *see* Nonstandardized tests
- novel, 422-23
- objective, *see* Objective tests
- planning, 140-47
- preparing, 147-55
- rearrangement, *see* Rearrangement tests
- reliability, *see* Reliability
- results, *see* Test results
- scale distinguished from, 24
- scores, *see* Scores
- scoring, *see* Scoring the tests
- selecting appropriate, 214-24
- simple-recall, *see* Simple-recall tests
- standardized, *see* Standardized tests
- time allowance for, 155-56
- trying out, *see* Trying out tests
- usability, *see* Usability
- validity, *see* Validity

Tests and Measurements in High School Instruction Ruch and Stoddard, 54

Tests in English Fundamentals Grammar 176-77

Tests of General Educational Development (GED), 422

Tests on Everyday Problems in Science Unit III, 177-78

Thematic Apperception Test (TAT), 421

Theory and Practice of Psychological Testing, Freeman, 55*Theory of Mental Tests*, Gulliksen, 55*Third Mental Measurements Yearbook* (1919), 55

Thorndike Handwriting Scale, 39

Thorndike-McCall Reading Test, 279, 295

Three-Year Study of Commission on Teacher Education, 374

Time allowance for test, 155-56

Trait variability, 352-53

Traxler Silent Reading Test, Word meaning, 183

Trying out the test, 155-58

answer keys, 158

conditions for, insuring normal, 155

scoring procedure, 156-58

scoring rules, 158

time allowance, 155-56

Two-way frequency table, 67-69

Typewriter graphs, 263

Ungrouped series, 61

Unit Scales of Attainment in Foods and Household Management, 183

University of Chicago High School System of reporting, 411-13

Usability

administration, ease, 127-28

application, ease, 129-30

cost, 130

essay examination, 195, 196

interpretation, ease, 129-30

meaning, 127

mechanical make-up, 130-31

quality of satisfactory measuring instrument, 127-31

recent tendencies, 57-58

scoring, ease, 128-29

Utilizing Human Talent, Davis, 55

Validity

achievement tests, 111-21

criticisms, 113-14

curricular vs statistical, 111-13

direct vs indirect methods, 115-17

item analysis, 117-19

standard tests, judging, 119-21

Tyler's suggestions, 114-15

coefficient, 101

considerations, general, 107-08

curricular, 101

statistical vs, 111-13

essay examination, 193, 196-97

intelligence tests, 108-11

individual vs group, 109-11

meaning of intelligence, 108

Terman criteria, 108-09

meaning, 107

objective tests, 164-67

quality of satisfactory measuring instrument, 107-21

- Validity (Cont.)
statistical, 101
curricular versus, 111-13
types 416-17
- Variability
concept of test data, 69, 71-72
human *see* Human variability
meaning 81
measures, 81-85
percentile, *D*, 81-85
quartile deviation, 81-83
range, 81
standard deviation, 83-85
vocabulary of, 72-73
- Variations, concomitant, *see* Concomitant
variation
- Vitiation, school, public information, 413
- Watson-Glaser Critical Thinking Appraisal, 422-23
- Wechsler-Bellevue Intelligence Scales, 55, 215
- Wechsler Intelligence Scale for Children (WISC), 55 418 422
- Wesley Test in Political Terms, 182
- Wholistic approach 420
- Woodworth Personal Data Sheet, 49
- World success criterion of intelligence, 109
- World War I 35-37, 48
- World War II, 37 55, 363
- λ -O Test, 49
- Z-scores, 85, 289-90